

Gibbs Reference Prior for Robust Gaussian Process Emulation.

Joseph Muré

March 22, 2017

Abstract

We propose an objective prior distribution on correlation kernel parameters for Simple Kriging models in the spirit of reference priors. Because it is proper and defined through its conditional densities, it and its associated posterior distribution lend themselves well to Gibbs sampling, thus making the full-Bayesian procedure tractable. Numerical examples show it has near-optimal frequentist performance in terms of prediction interval coverage.

1 Introduction

Gaussian Processes are widely used to model the spatial distribution of some real-valued quantity when said quantity is only observed at a few locations. This emulation technique is a convenient way to represent the uncertainty of the value of the quantity at unobserved points [Santner et al., 2003]. In this work, we focus on stationary Gaussian Processes with null mean function, a framework that is referred to as “Simple Kriging” in the geostatistical literature [Journel and Huijbregts, 1978] and is also frequently used in the context of computer experiments and machine learning [Rasmussen and Williams, 2006]. The exact probability distribution of a Gaussian Process depends not only on its mean function (supposed here to be null), but also on a variance parameter and a correlation function (also known as “correlation kernel”) which itself depends on parameters.

We propose an “objective” Jeffreys-like prior distribution on these parameters, which we call “Gibbs reference prior”.

The need for a prior distribution on the parameters of Kriging models arises from the lack of robustness of the Maximum Likelihood Estimator (MLE) in dealing with parameters of correlation kernels. Indeed, what makes estimating them “notoriously difficult”, as Kennedy and O’Hagan [2001] put it, is that the likelihood function may often be quite flat [Li and Sudjianto, 2005]. To tackle this problem, one may stabilize the MLE by adding a nugget to the covariance kernel, namely adding a covariance component concentrated on the diagonal. However, as was noted by Andrianakis and Challenor [2012], “the presence of a nugget is equivalent to the assumption that the simulator contains some variability that is not explainable by its inputs”. Alternatively, Li and Sudjianto [2005] proposed penalizing the likelihood function, which may also be interpreted as using a prior distribution and then choosing the Maximum A Posteriori (MAP) estimate instead of the MLE. Of course, using a full-Bayesian approach obviates the problem of robustness of the estimator of the parameters, as one may simply use the integrated predictive distribution.

Whether one wishes to use a prior distribution as a penalizing function for the likelihood or to deploy the whole Bayesian machinery, one often faces the problem of a lack of *a priori* information. This is where Objective Bayes, which was first introduced in this context by Berger et al. [2001], is helpful. The authors’ work on deriving the reference prior in this context and establishing posterior propriety was then successively extended by Paulo [2005], Ren et al. [2013] and Gu [2016]. Ren et al. [2012] and Kazianka and Pilz [2012] also derived and studied the reference prior in the case of an added nugget effect. However, the above cited works all make a restrictive assumption in order to guarantee posterior propriety, which essentially implies that for any twice differentiable correlation kernel, the number of observation points must not exceed the spatial dimension by more than 2 (see Appendix A for details). Standard covariance kernels such as the Matérn covariance function with smoothness parameter $\nu > 1$ thus cannot be used. This, along with the wish to make the full-Bayesian process tractable in practice, led us to consider a different but similar “objective” prior distribution. As it is defined through conditional densities and thus is well suited to Gibbs sampling, we call it the “Gibbs reference prior”.

Because we do not use the exact reference prior, we need to establish what we mean with the notion of an “objective” prior distribution. The following rule-of-thumb lays out our intent. An “objective” prior distribution should fulfill two requirements :

1. It should not require any user input, that is, it should depend on no user-specified metaparameter.
2. It should rely on the information one may expect from the observations given by the design set. For instance, like the reference prior, the Gibbs reference prior relies on the Fisher information.

It should be noted that the second requirement implies an “objective” prior distribution necessarily depends on the design set, because where we observe the Gaussian Process bears heavily on the amount of information we may expect from the observations. In particular, it means the Lebesgue measure cannot be used as an “objective” prior distribution. Besides, this is consistent with the dependence of the asymptotic performance of the MLE on the type of design set [Bachoc, 2014].

We offer theoretical guarantees of posterior propriety in the case where a Matérn class covariance kernel [Matérn, 1986] [Handcock and Stein, 1993] with known and finite smoothness parameter is used. For a thorough description of Matérn class covariance kernels and their properties, see Stein [1999] or Bachoc [2013]. Moreover, we provide a framework which could be used to prove this result for other classes of correlation kernels. Section 2 describes this framework, while section 3 shows how to apply it to Matérn class correlation kernels. Section 4 indicates how to sample from the posterior distribution on parameters and contrasts two ways of using it, namely the full-Bayesian method and the MAP plug-in approach.

The MAP plug-in approach has the intrinsic disadvantage of requiring parameter estimation. As its effectiveness depends on the quality of the estimation, it makes sense to compare its performance with that of the MLE. Section 5 shows that the MAP estimator improves inference robustness significantly when compared to the MLE.

Beyond parameter inference, what matters to us is how well we are able to account for uncertainty on values of the Gaussian Process at unobserved points. Section 6 shows that predictive intervals at unobserved points produced by the full-Bayesian method have effective coverage close to their theoretical level, while

predictive intervals produced by estimator plug-in methods (MAP and *a fortiori* MLE) have substantially lower effective coverage.

2 Objective prior distribution

We assume that the random field of interest $\{Y(\mathbf{x}), \mathbf{x} \in D\}$, D being a bounded subset of \mathbb{R}^r , is Gaussian, with zero mean (or known mean) and with covariance of the form $\text{Cov}(Y(\mathbf{x}), Y(\mathbf{x}')) = \sigma^2 K_{\boldsymbol{\theta}}(\mathbf{x} - \mathbf{x}')$. σ^2 thus denotes the variance of the Gaussian Process and $\boldsymbol{\theta}$, hereafter named the “vector of correlation lengths”, is the vector of scaling parameters used by the chosen class of correlation kernels $K_{\boldsymbol{\theta}}$. Although the results of the first sections are true for any class of correlation kernels, we believe the reader would do well to keep in mind we will later focus on the case where it represents the vector of correlation lengths of Matérn class correlation kernels (but not their smoothness parameter, which will be assumed to be known).

\mathbf{y} denotes the $n \times 1$ vector of observations made at the points of the design set and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ the correlation matrix of the Gaussian vector of which \mathbf{y} is a realization. When applied to a matrix, $|\cdot|$ refers to its determinant.

With these notations, the distribution of \mathbf{y} given σ^2 and $\boldsymbol{\theta}$ is $\mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$ and the likelihood of parameters σ^2 and $\boldsymbol{\theta}$ is :

$$L(\mathbf{y} \mid \sigma^2, \boldsymbol{\theta}) = \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} |\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \mathbf{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{y} \right\}. \quad (1)$$

The Gibbs reference prior is, as indicated by its name, based on Bernardo’s reference prior [Bernardo, 1979] [Berger and Bernardo, 1992]. Under a few regularity conditions [Clarke and Barron, 1994], the reference prior as defined in Berger et al. [2009] coincides with the Jeffreys-rule prior. However, the Jeffreys-rule prior has known shortcomings in cases with several unknown parameters.

For example, consider the case of n independent real-valued observations y_1, \dots, y_n of the model $\mathcal{N}(\mu, \sigma^2)$ [Robert et al., 2009]. If μ is known but σ^2 is not, the Jeffreys-rule prior on σ^2 is $(\sigma^2)^{-1}$. Then the posterior distribution of $\sum_{i=1}^n (y_i - \bar{y})^2 / \sigma^2$ (*i.e.* its distribution knowing all y_i ($1 \leq i \leq n$) and μ) is the chi-squared distribution with n degrees of freedom.

If both μ and σ^2 are unknown, the Jeffreys-rule joint prior distribution on μ and σ^2 is $(\sigma^2)^{-3/2}$. Let us denote $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Then the posterior distribution of $\sum_{i=1}^n (y_i - \bar{y})^2 / \sigma^2$ (*i.e.* its distribution knowing all y_i ($1 \leq i \leq n$) but not μ) is *also* the chi-squared distribution with n degrees of freedom. In other words, using the Jeffreys-rule prior in both situations implies that by simply substituting the empirical mean \bar{y} to the actual mean μ , we are able to reach the same state of knowledge about σ^2 as if we actually knew μ . Alternatively, one may use the independence Jeffreys prior distribution, which is the joint prior distribution on μ and σ^2 obtained by taking the product of the Jeffreys-rule prior on μ when σ^2 is known (flat prior) and of the Jeffreys-rule prior on σ^2 when μ is known, thus yielding the joint prior distribution $(\sigma^2)^{-1}$. Then the posterior distribution of $\sum_{i=1}^n (y_i - \bar{y})^2 / \sigma^2$ is the chi-squared distribution with $n-1$ degrees of freedom, which acknowledges the loss of information on σ^2 when μ is unknown. This suggests that separating parameters may improve performance of Jeffreys priors in multidimensional cases.

The reference prior algorithm was designed to overcome such shortcomings [Berger and Bernardo, 1992]. It first requires the user to specify an ordering on the parameters. Then, based on this ordering and for each

parameter, the reference prior (*i.e.* the Jeffreys-rule prior in regular cases) on the parameter is computed, the parameter is integrated out of the likelihood function and the likelihood function is replaced by the new, “integrated” version.

Note that in the example, whatever the chosen ordering of μ and σ^2 , the reference prior coincides with the independence Jeffreys prior.

To apply the reference prior algorithm, we first need to decide on an ordering of σ^2 and θ . Knowledge of θ implies knowledge of correlation matrix Σ_θ , which may then be used to decorrelate the vector of observations \mathbf{y} . For any matrix $\sqrt{\Sigma_\theta}$ such that $\Sigma_\theta = \sqrt{\Sigma_\theta} \sqrt{\Sigma_\theta}^\top$ (the Cholesky decomposition for instance), if σ^2 and θ are known, then $\mathbf{z} := \sqrt{\Sigma_\theta}^{-1} \mathbf{y}$ follows the multivariate normal distribution $\mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$. This means that if we assume only θ is known, the inference problem is reduced to the aforementioned case of n independent observations of a normal model with known mean but unknown variance. Conversely, knowledge of σ^2 may be used to rescale the vector of observations \mathbf{y} : if σ^2 and θ are known, $\mathbf{z}' = (\sigma^2)^{-1/2} \mathbf{y}$ follows the multivariate normal distribution $\mathcal{N}(\mathbf{0}_n, \Sigma_\theta)$, but this inference problem is nearly as difficult as the original one. This dissymmetry motivates our choice to first derive the reference prior on σ^2 when θ is known, and then the marginal prior on θ .

The Fisher information of the model on σ^2 assuming θ to be known is $\frac{\sqrt{n}}{2\sigma^2}$. Thus we define the prior distribution on σ^2 knowing θ as $\pi(\sigma^2 | \theta) \propto \frac{1}{\sigma^2}$.

Let us average the likelihood with respect to this distribution :

$$\begin{aligned} L^1(\mathbf{y} | \theta) &\propto \int_0^\infty L(\mathbf{y} | \sigma^2, \theta) \pi(\sigma^2 | \theta) d(\sigma^2) \\ &\propto (2\pi)^{-\frac{n}{2}} |\Sigma_\theta|^{-\frac{1}{2}} \left(\frac{\mathbf{y}^\top \Sigma_\theta^{-1} \mathbf{y}}{2} \right)^{-\frac{n}{2}} \Gamma\left(\frac{n}{2}\right) \propto (\mathbf{y}^\top \Sigma_\theta^{-1} \mathbf{y})^{-\frac{n}{2}}. \end{aligned} \quad (2)$$

Assuming correlation length θ to be known, $\mathbf{y}^\top \Sigma_\theta^{-1} \mathbf{y} / n$ is the MLE of σ^2 . Plugging this estimator into the likelihood of the model (1), we get an expression that is proportional to integrated likelihood (2) : $L(\mathbf{y} | \sigma^2 = \frac{1}{n} \mathbf{y}^\top \Sigma_\theta^{-1} \mathbf{y}, \theta) \propto L^1(\mathbf{y} | \theta)$. Thus, using the MLE for σ^2 is equivalent to integrating the likelihood of the model (1) over the prior distribution on σ^2 .

We suppose from now on that $L^1(\mathbf{y} | \theta)$ remains bounded as θ varies. This assumption is not very restrictive (see Appendix B) and it ensures any proper prior distribution on θ yields a proper posterior distribution.

2.1 One-dimensional θ

In this section, we assume that θ is a scalar parameter. To emphasize this, we denote it θ .

The next step in the reference prior algorithm requires us to compute the reference prior on θ with respect to the integrated likelihood $L^1(\cdot | \theta)$. Defining the reference prior as the Jeffreys-rule prior, this requires us to compute the Fisher information on θ under the measure μ_θ with density $L^1(\cdot | \theta)$ with respect to the Lebesgue measure on \mathbb{R}^n . However, as μ_θ is no finite measure, this quantity has no obvious definition.

If μ_θ were a probability measure, the Fisher information on θ would be the standard deviation under $\mu_\theta(\mathbf{y})$ of $\partial_\theta [\log L^1(\mathbf{y} | \theta)]$, which is only defined up to an additive constant C_θ .

$$\partial_\theta [\log L^1(\mathbf{y} | \theta)] = -\frac{n}{2} \frac{\mathbf{y}^\top \partial_\theta (\boldsymbol{\Sigma}_\theta^{-1}) \mathbf{y}}{\mathbf{y}^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{y}} + C_\theta. \quad (3)$$

As C_θ does not depend on \mathbf{y} , we may safely define its “variance under $\mu_\theta(\mathbf{y})$ ” as 0. Thus, if we can define the “variance under $\mu_\theta(\mathbf{y})$ ” of $-n/2 \mathbf{y}^\top \partial_\theta (\boldsymbol{\Sigma}_\theta^{-1}) \mathbf{y} / \mathbf{y}^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{y}$, we will be able to take it to be the “variance under $\mu_\theta(\mathbf{y})$ ” of $\partial_\theta [\log L^1(\mathbf{y} | \theta)]$.

Define a matrix $\sqrt{\boldsymbol{\Sigma}_\theta}$ such that $\boldsymbol{\Sigma}_\theta = \sqrt{\boldsymbol{\Sigma}_\theta} \sqrt{\boldsymbol{\Sigma}_\theta}^\top$. Then $\sqrt{\boldsymbol{\Sigma}_\theta}^{-1}$ is such that $\boldsymbol{\Sigma}_\theta^{-1} = (\sqrt{\boldsymbol{\Sigma}_\theta}^{-1})^\top \sqrt{\boldsymbol{\Sigma}_\theta}^{-1}$. Now let $\sqrt{\boldsymbol{\Sigma}_\theta}^{-1} * \mu_\theta$ be the pushforward measure of μ_θ resulting from premultiplication by $\sqrt{\boldsymbol{\Sigma}_\theta}^{-1}$. Its density with respect to Lebesgue measure λ is given by :

$$\frac{d\sqrt{\boldsymbol{\Sigma}_\theta}^{-1} * \mu_\theta(\mathbf{w})}{d\lambda} = \left| \sqrt{\boldsymbol{\Sigma}_\theta} \right| L^1 \left(\sqrt{\boldsymbol{\Sigma}_\theta} \mathbf{w} | \theta \right) \propto (\mathbf{w}^\top \mathbf{w})^{-\frac{n}{2}} \quad (4)$$

As the singleton $\{(0, \dots, 0)\}$ is negligible with respect to the measure $\sqrt{\boldsymbol{\Sigma}_\theta}^{-1} * \mu_\theta$, we redefine henceforth $\sqrt{\boldsymbol{\Sigma}_\theta}^{-1} * \mu_\theta$ as being a measure on $\mathbb{R}^n \setminus \{(0, \dots, 0)\}$. Now, let f be the bijection $\mathbb{R}^n \setminus \{(0, \dots, 0)\} \rightarrow S^{n-1} \times (0, +\infty)$ such that $f(\mathbf{w}) = \left(\frac{\mathbf{w}}{\sqrt{\mathbf{w}^\top \mathbf{w}}}, \sqrt{\mathbf{w}^\top \mathbf{w}} \right)$. The pushforward measure $f * \sqrt{\boldsymbol{\Sigma}_\theta}^{-1} * \mu_\theta$ can be written as $U \otimes \rho$, where U is the uniform probability distribution on S^{n-1} and ρ is the measure whose Radon-Nikodym derivative with respect to Lebesgue measure is $\frac{d\rho}{d\lambda}(r) \propto r^{-n} A(r) \propto r^{-1}$, where $A(r)$ is the area of the sphere with radius r .

Hence it seems reasonable to define $\frac{\mathbf{w}}{\sqrt{\mathbf{w}^\top \mathbf{w}}}$ as following “under $\sqrt{\boldsymbol{\Sigma}_\theta}^{-1} * \mu_\theta(\mathbf{w})$ ” the uniform distribution on the unit sphere S^{n-1} . From there, we get that $\frac{\sqrt{\boldsymbol{\Sigma}_\theta}^{-1} \mathbf{y}}{\sqrt{\mathbf{y}^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{y}}}$ follows “under $\mu_\theta(\mathbf{y})$ ” the uniform distribution on the unit sphere S^{n-1} .

We are now able to define the “variance (resp. standard deviation) under $\mu_\theta(\mathbf{y})$ ” of $\partial_\theta [\log L^1(\mathbf{y} | \theta)]$. It is $\text{Var}[\mathbf{U}^\top \mathbf{M}_\theta^\Sigma \mathbf{U}]$ (resp. $\text{Var}[\mathbf{U}^\top \mathbf{M}_\theta^\Sigma \mathbf{U}]^{1/2}$), where \mathbf{U} is a random variable following the uniform distribution on S^{n-1} and

$$\mathbf{M}_\theta^\Sigma := \left(\sqrt{\boldsymbol{\Sigma}_\theta} \right)^\top \partial_\theta (\boldsymbol{\Sigma}_\theta^{-1}) \sqrt{\boldsymbol{\Sigma}_\theta} \quad (5)$$

Proposition 1. *The above defined “standard deviation under $\mu_\theta(\mathbf{y})$ ” of $\partial_\theta [\log L^1(\mathbf{y} | \theta)]$, which we take as being proportional to our prior distribution $\pi(\theta)$, is*

$$\pi(\theta) \propto \sqrt{\text{Tr} \left[\left(\frac{\partial}{\partial \theta} (\boldsymbol{\Sigma}_\theta) \boldsymbol{\Sigma}_\theta^{-1} \right)^2 \right] - \frac{1}{n} \text{Tr} \left[\frac{\partial}{\partial \theta} (\boldsymbol{\Sigma}_\theta) \boldsymbol{\Sigma}_\theta^{-1} \right]^2}. \quad (6)$$

The proof of proposition 1 is given in Appendix C.

Let us note that the full prior distribution $\pi(\sigma^2, \theta) \propto \frac{1}{\sigma^2} \pi(\theta)$ is the same as the Jeffreys-rule prior distribution computed from the original likelihood function $L(\mathbf{y} | \sigma^2, \theta)$ (which is also the reference prior distribution given by theorem 2 of Berger et al. [2001] when taken in the Simple Kriging framework).

2.2 Multi-dimensional θ

In the case where θ has dimension $r > 1$, there is no natural way of setting up a hierarchy between correlation lengths. It is thus tempting to group them. This requires us to compute the Fisher information matrix $\mathcal{I}(\theta)$ related to the correlation lengths.

Proposition 1 yields the diagonal coefficients of this matrix. Non-diagonal coefficients may be derived using the polarization formula. For every integer i and j between 1 and r ,

$$[\mathcal{I}(\theta)]_{ij} \propto \text{Tr} \left[\left(\frac{\partial}{\partial \theta_i} (\Sigma_\theta) \Sigma_\theta^{-1} \right) \left(\frac{\partial}{\partial \theta_j} (\Sigma_\theta) \Sigma_\theta^{-1} \right) \right] - \frac{1}{n} \text{Tr} \left[\frac{\partial}{\partial \theta_i} (\Sigma_\theta) \Sigma_\theta^{-1} \right] \text{Tr} \left[\frac{\partial}{\partial \theta_j} (\Sigma_\theta) \Sigma_\theta^{-1} \right] \quad (7)$$

This reference prior, which can be written $\pi_R(\sigma^2, \theta) \propto \frac{1}{\sigma^2} \sqrt{|\mathcal{I}(\theta)|}$, does not coincide with that of Paulo [2005], nor with any reference prior given by Ren et al. [2013]. It is a different extension of Berger et al. [2001]'s prior to the case of a multidimensional θ (although it is also a restriction, because this prior distribution only fits the Simple Kriging framework whereas Berger et al. [2001]'s, Paulo [2005]'s and Ren et al. [2013]'s all fit the Universal Kriging framework).

However, this method has the disadvantage of requiring the use of a multidimensional Jeffreys-rule prior distribution, which may show the sort of undesirable behavior discussed at the beginning of this section. Alternatively, we could draw inspiration from the one-dimensional case in the following way. Suppose that we know every entry of θ except one, θ_i . Then, according to equation (6), the prior density on θ_i knowing all entries θ_j ($j \neq i$) would be

$$\pi(\theta_i \mid \theta_j \forall j \neq i) \propto \sqrt{\text{Tr} \left[\left(\frac{\partial}{\partial \theta_i} (\Sigma_\theta) \Sigma_\theta^{-1} \right)^2 \right] - \frac{1}{n} \text{Tr} \left[\frac{\partial}{\partial \theta_i} (\Sigma_\theta) \Sigma_\theta^{-1} \right]^2}. \quad (8)$$

The idea is to regard all conditional densities $\pi(\theta_i \mid \theta_j \forall j \neq i)$ ($1 \leq i \leq r$) as implicitly defining a joint prior density on all entries θ_i ($1 \leq i \leq r$). This has several advantages. First, it allows us to treat all θ_i s separately instead of using a multidimensional Jeffreys-rule prior. Second, it sidesteps the need to specify an ordering on θ_i s, which would be arbitrary in the absence of additional prior knowledge. Third, it enables Gibbs sampling (or rather Metropolis within Gibbs sampling), which is less computationally intensive than multidimensional Metropolis sampling.

What we now need to show is that, for certain classes of covariance kernels, such a joint prior density exists and is uniquely defined by the conditional densities $\pi(\theta_i \mid \theta_j \forall j \neq i)$.

In this paper, we prove this result for all correlation kernels that satisfy the following assumption. As shown in Section 3, Matérn kernels are among them.

Let us at this point introduce new notations : $\forall 1 \leq i \leq r$, we define

$$\underline{\theta}_r := \min(\theta_1, \theta_2, \dots, \theta_r) \text{ and } \underline{\theta}_{r,i} := \min(\theta_j \mid 1 \leq j \leq r, j \neq i). \quad (9)$$

Assumption 1. *There exist real numbers $a \geq 0$ and $b > 1$ verifying the following statement :*

For every integer i between 1 and r , there exist nonnegative real numbers T_i , $M_{i,1}$, $M_{i,2}$, $M_{i,3}$, $M_{i,4}$, $M_{i,5}$ and $M_{i,6}$ such that

1. $\forall \boldsymbol{\theta} \in (\mathbb{R}_+)^r$, $\|\frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\| \leq M_{i,1}$.
2. $\forall \boldsymbol{\theta} \in (\mathbb{R}_+)^r$, $\|\theta_i^b \frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\| \leq M_{i,2}$.
3. $\forall \boldsymbol{\theta} \in (\mathbb{R}_+)^r$, $\|\underline{\theta}_{r,i}^{-a} \frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\| \leq M_{i,3}$.
4. $\forall \boldsymbol{\theta} \in (\mathbb{R}_+)^r$, $\|\underline{\theta}_{r,i}^{-a} \theta_i^b \frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\| \leq M_{i,4}$.
5. $\forall \boldsymbol{\theta} \in (\mathbb{R}_+)^r$ such that $\underline{\theta}_{r,i} \leq T_i$, $\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\| \leq M_{i,5}$.
6. $\forall \boldsymbol{\theta} \in (\mathbb{R}_+)^r$ such that $\underline{\theta}_{r,i} \geq T_i$, $\|\underline{\theta}_{r,i}^{-a} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\| \leq M_{i,6}$.

Let us briefly discuss the majorations of Assumption 1.

- Majoration 1 is only required to make sure $\|\frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\|$ cannot be arbitrarily great when θ_i is near 0, which would compromise the propriety of the conditional distributions (8).
- Majoration 2 makes $\|\frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\|$ decrease “fast” when θ_i increases to infinity. This speed determines the size of the tails of the conditional prior distributions (8). The greater b is, the thinner their tails. If b is not taken greater than 1, we cannot guarantee their being proper distributions.
- Majoration 5 is rather straightforward : provided one of the “known” entries of $\boldsymbol{\theta}$ is small enough, it allows us to control $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$ and prevent it from interfering with $\frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ ’s behavior, which is important to ensure the propriety of the conditional distributions (8).
- Majoration 6 is in practice the most difficult to prove. If all entries of $\boldsymbol{\theta}$ go simultaneously to infinity, it is natural that $\|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\|$ should also go to infinity, but we need to control how fast it goes there. The idea is that as far as conditional density $\pi(\theta_i \mid \theta_j \forall j \neq i)$ is concerned, $\underline{\theta}_{r,i}$ is a multiplicative constant, which means that we could divide the conditional density by $\underline{\theta}_{r,i}^a$ to make it more manageable without changing the conditional distribution. However, this change increases the value of the density for small $\underline{\theta}_{r,i}$, which requires us to “check again” majorations 1 and 2 for small $\underline{\theta}_{r,i}$, thus yielding majorations 3 and 4 respectively. In summary, a needs to be great enough to ensure majoration 6, while also being small enough to not violate majorations 3 and 4.

Let us define

$$d(\theta_i \mid \theta_j \forall j \neq i) := \underline{\theta}_{r,i}^{-a} \sqrt{[\mathcal{I}(\boldsymbol{\theta})]_{ii}}. \quad (10)$$

Obviously, $d(\theta_i \mid \theta_j \forall j \neq i) \propto \pi(\theta_i \mid \theta_j \forall j \neq i)$.

Lemma 2. *If Assumption 1 holds, then for every integer i such that $1 \leq i \leq r$, there exists a nonnegative real number M_i such that $\forall \boldsymbol{\theta} \in (\mathbb{R}_+)^r$, $d(\theta_i \mid \theta_j \forall j \neq i) \leq M_i \min\left(1, \frac{1}{\theta_i^b}\right)$.*

Proof. First, it follows from the definition of $d(\theta_i \mid \theta_j \forall j \neq i)$ that

$$d(\theta_i \mid \theta_j \forall j \neq i)^2 \leq \text{Tr} \left(\left(\underline{\theta}_{r,i}^{-a} \left(\frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \right) \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \right)^2 \right). \quad (11)$$

Then, noting $\|\cdot\|$ the Frobenius norm on matrices, it follows from the Cauchy-Schwarz inequality $\text{Tr}(\mathbf{A}^\top \mathbf{B}) \leq \|\mathbf{A}\| \|\mathbf{B}\|$ (with $\mathbf{B} := \underline{\theta}_{r,i}^{-a} \left(\frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_\theta \right) \boldsymbol{\Sigma}_\theta^{-1}$ and $\mathbf{A} := \mathbf{B}^\top$) that

$$d(\theta_i \mid \theta_j \forall j \neq i) \leq \left\| \underline{\theta}_{r,i}^{-a} \left(\frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_\theta \right) \boldsymbol{\Sigma}_\theta^{-1} \right\|. \quad (12)$$

The conclusion then follows from Assumption 1 :

- when $\underline{\theta}_{r,i} \leq T_i$, after observing that $\left\| \underline{\theta}_{r,i}^{-a} \left(\frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_\theta \right) \boldsymbol{\Sigma}_\theta^{-1} \right\| \leq \left\| \underline{\theta}_{r,i}^{-a} \left(\frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_\theta \right) \right\| \|\boldsymbol{\Sigma}_\theta^{-1}\|$, it is yielded by combining majorations 3, 4 and 5 ;
- when $\underline{\theta}_{r,i} \geq T_i$, after observing that $\left\| \underline{\theta}_{r,i}^{-a} \left(\frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_\theta \right) \boldsymbol{\Sigma}_\theta^{-1} \right\| \leq \left\| \left(\frac{\partial}{\partial \theta_i} \boldsymbol{\Sigma}_\theta \right) \right\| \left\| \underline{\theta}_{r,i}^{-a} \boldsymbol{\Sigma}_\theta^{-1} \right\|$, it is yielded by combining majorations 1, 2 and 6.

□

It follows from lemma 2 that, under Assumption 1, for every integer i such that $1 \leq i \leq r$, whatever the value of the vector $(\theta_j)_{j \neq i}$, the function $d(\theta_i = \cdot \mid \theta_j \forall j \neq i)$ is the density of a finite measure on \mathbb{R}_+ .

Theorem 3. *If $r \geq 2$ and Assumption 1 holds, then there exists a unique probability density $\pi(\boldsymbol{\theta})$ on $(\mathbb{R}_+)^r$ with respect to Lebesgue measure such that, for every integer verifying $1 \leq i \leq r$, whatever the value of the vector $(\theta_j)_{j \neq i}$, the conditional density $\pi(\theta_i \mid \theta_j \forall j \neq i)$ is proportional to $d(\theta_i \mid \theta_j \forall j \neq i)$.*

$\pi(\boldsymbol{\theta})$ will be referred to as the ‘‘Gibbs reference prior density on $\boldsymbol{\theta}$ ’’. The following result allows control of its tail rates.

Theorem 4. *If $r \geq 2$ and Assumption 1 holds, then for a certain positive number M , the Gibbs reference prior density $\pi(\boldsymbol{\theta})$ verifies*

$$\forall \boldsymbol{\theta} \in (0, +\infty)^r, \quad \pi(\boldsymbol{\theta}) \leq M \prod_{i=1}^r \theta_i^{-b}. \quad (13)$$

The proofs of Theorem 3 and Theorem 4 are given in Appendix C.

3 Example : Matérn kernels

This section aims to give an example of a commonly used covariance kernel family which satisfies Assumption 1 under a mild assumption on the design set. To express this assumption in a convenient way, we use the following definition. A set has coordinate-distinct points if for any distinct points in the set \mathbf{x} and \mathbf{x}' , every component of vector $\mathbf{x} - \mathbf{x}'$ differs from 0.

Assumption 2. *The design set has coordinate-distinct points.*

Randomly sampled (through Latin Hypercube Sampling for example) design sets almost surely have coordinate-distinct points. Cartesian product design sets, however, do not satisfy Assumption 2.

In this work, we follow the definition of Matérn kernels given in Bachoc [2013] (see next paragraph). We use the following convention for the Fourier transform : the Fourier transform \widehat{g} of a function $g : \mathbb{R}^r \rightarrow \mathbb{R}$ verifies $g(\mathbf{x}) = \int_{\mathbb{R}^r} \widehat{g}(\boldsymbol{\omega}) e^{i\langle \boldsymbol{\omega} | \mathbf{x} \rangle} d\boldsymbol{\omega}$ and $\widehat{g}(\boldsymbol{\omega}) = (2\pi)^{-r} \int_{\mathbb{R}^r} g(\mathbf{x}) e^{-i\langle \boldsymbol{\omega} | \mathbf{x} \rangle} d\mathbf{x}$.

Now let us set a few notations :

- B_ν is the modified Bessel function of second kind with parameter ν ;
- $K_{r,\nu}$ is the r -dimensional Matérn isotropic covariance kernel with variance 1, correlation length 1 and smoothness $\nu \in (0, +\infty)$ and $\widehat{K}_{r,\nu}$ is its Fourier transform :

$$- \forall \mathbf{x} \in \mathbb{R}^r, \quad K_{r,\nu}(\mathbf{x}) = \frac{1}{\Gamma(\nu)2^{\nu-1}} (2\sqrt{\nu}\|\mathbf{x}\|)^\nu B_\nu(2\sqrt{\nu}\|\mathbf{x}\|) ; \quad (14)$$

$$- \forall \boldsymbol{\omega} \in \mathbb{R}^r, \quad \widehat{K}_{r,\nu}(\boldsymbol{\omega}) = \frac{M_r(\nu)}{(\|\boldsymbol{\omega}\|^2 + 4\nu)^{\nu+\frac{r}{2}}} \text{ with } M_r(\nu) = \frac{\Gamma(\nu + \frac{r}{2})(2\sqrt{\nu})^{2\nu}}{\pi^{\frac{r}{2}}\Gamma(\nu)}. \quad (15)$$

- $K_{r,\nu}^{tens}$ is the r -dimensional Matérn tensorized covariance kernel with variance 1, correlation length 1 and smoothness $\nu \in \mathbb{R}_+$ and $\widehat{K}_{r,\nu}^{tens}$ is its Fourier transform :

$$- \forall \mathbf{x} \in \mathbb{R}^r, \quad K_{r,\nu}^{tens}(\mathbf{x}) = \prod_{j=1}^r K_{1,\nu}(\mathbf{x}_j) ; \quad (16)$$

$$- \forall \boldsymbol{\omega} \in \mathbb{R}^r, \quad \widehat{K}_{r,\nu}^{tens}(\boldsymbol{\omega}) = \prod_{j=1}^r \widehat{K}_{1,\nu}(\boldsymbol{\omega}_j). \quad (17)$$

- let us adopt the following convention : if $\mathbf{t} \in \mathbb{R}^r$, $\frac{\mathbf{t}}{\boldsymbol{\theta}} = \left(\frac{t_1}{\theta_1}, \dots, \frac{t_r}{\theta_r}\right)$.

We define the Matérn geometric anisotropic covariance kernel with variance parameter σ^2 , correlation lengths $\boldsymbol{\theta}$ and smoothness ν as the function $\mathbf{x} \mapsto \sigma^2 K_{r,\nu}\left(\frac{\mathbf{x}}{\boldsymbol{\theta}}\right)$.

Similarly, we define the Matérn tensorized covariance kernel with variance parameter σ^2 , correlation lengths $\boldsymbol{\theta}$ and smoothness ν as the function $\mathbf{x} \mapsto \sigma^2 K_{r,\nu}^{tens}\left(\frac{\mathbf{x}}{\boldsymbol{\theta}}\right)$.

Proposition 5. *For design sets with coordinate-distinct points in a r -dimensional domain D ($r \geq 2$) and Matérn tensorized or anisotropic geometric covariance kernels, Assumption 1 is true with any $a \geq 2\nu$ and $1 < b < 1 + 2\min(1, \nu)$. In fact, whenever $\nu \neq 1$, it is true with any $a \geq 2\nu$ and $1 < b \leq 1 + 2\min(1, \nu)$.*

The proof of Proposition 5 is given in Appendix D.

This result means that as long as a design set with coordinate-distinct points is used, the Gibbs reference prior distribution on a multidimensional $\boldsymbol{\theta}$ exists and is proper when using both kinds of Matérn anisotropic covariance kernels.

4 Using the posterior distribution

4.1 Sampling of the posterior distribution

Because we can only explicitly access the prior distribution on $\boldsymbol{\theta}$ through its conditional distributions (up to a multiplicative constant), we will also only be able to access the posterior distribution this way :

$$\pi(\theta_i \mid \mathbf{y}, \theta_j \forall j \neq i) = L^1(\mathbf{y} \mid \boldsymbol{\theta})\pi(\theta_i \mid \theta_j \forall j \neq i). \quad (18)$$

Therefore, Gibbs sampling allows us to sample the posterior distribution on $\boldsymbol{\theta}$. As $\pi(\theta_i | \mathbf{y}, \theta_j \forall j \neq i)$ is nonstandard and only known up to a multiplicative constant, we use the Metropolis algorithm. Specifically, we use in the examples covered in sections 5 and 6 (except the higher-dimensional case studied in subsection 6.3) a normal instrumental density with standard deviation 0.2 and a 20-step burn-in period : in these cases, $\boldsymbol{\theta}$ is the vector of correlation lengths of the Matérn family of covariance functions and the spatial domain D is the unit cube $[0, 1]^3$. This leads to an acceptance rate of 0.65 and yields samples whose kernel-reconstructed conditional densities match $\pi(\theta_i | \mathbf{y}, \theta_j \forall j \neq i)$ as defined by (18).

Figure 1 shows the marginal Gibbs reference posterior distribution in a 3-dimensional case where the true correlation lengths vector is 0.4-0.8-0.2 and the true variance parameter 1. The design set contains 30 observation points. The represented density was estimated from a sample of 1000 points.

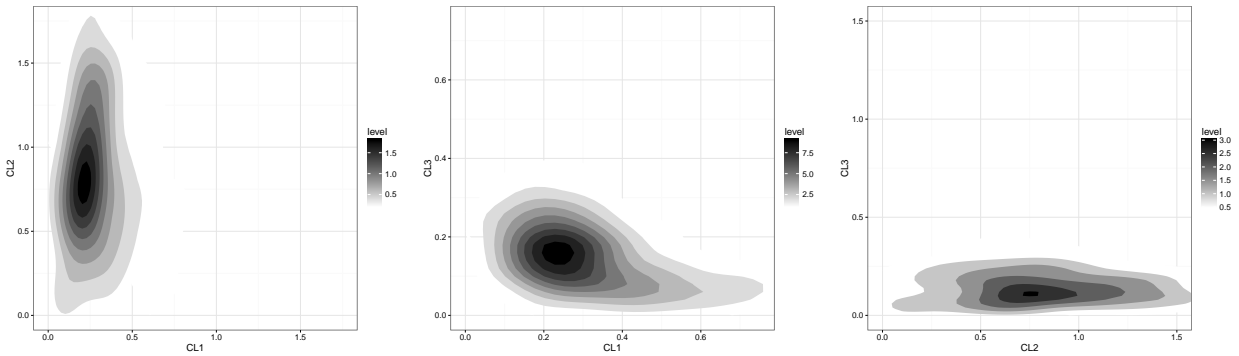


Figure 1: Marginal posterior densities of the joint first and second (left), first and third (middle), and second and third (right) correlation lengths. The true correlation lengths are 0.4 – 0.8 – 0.2. The design set contains 30 observations points that were uniformly sampled inside the unit cube.

4.2 From the posterior distribution to the predictive distribution

The ultimate goal of emulators is to supply a probability distribution on the value \mathbf{y}_0 of the quantity of interest at any unobserved set of points \mathbf{x}_0 .

The first step is to derive the joint posterior probability distribution on σ^2 and $\boldsymbol{\theta}$

$$\pi(\sigma^2, \boldsymbol{\theta} | \mathbf{y}) = \pi(\sigma^2 | \mathbf{y}, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} | \mathbf{y}). \quad (19)$$

This in turns requires us to compute $\pi(\sigma^2 | \mathbf{y}, \boldsymbol{\theta})$. Using Bayes' rule, this should be

$$\pi(\sigma^2 | \mathbf{y}, \boldsymbol{\theta}) = \frac{L(\mathbf{y} | \sigma^2, \boldsymbol{\theta}) \pi(\sigma^2 | \boldsymbol{\theta})}{L^1(\mathbf{y} | \boldsymbol{\theta})}. \quad (20)$$

However, neither $\pi(\sigma^2 | \boldsymbol{\theta}) = \pi(\sigma^2)$ nor $L^1(\mathbf{y} | \boldsymbol{\theta})$ are probability densities. In fact, they are only defined up to multiplicative constants. But equation (2) shows that both share the same multiplicative constant so their quotient does not depend on this constant. Moreover,

$$\int \pi(\sigma^2 | \mathbf{y}, \boldsymbol{\theta}) d\sigma^2 = \frac{\int L(\mathbf{y} | \sigma^2, \boldsymbol{\theta}) \pi(\sigma^2 | \boldsymbol{\theta}) d\sigma^2}{L^1(\mathbf{y} | \boldsymbol{\theta})} = \frac{L^1(\mathbf{y} | \boldsymbol{\theta})}{L^1(\mathbf{y} | \boldsymbol{\theta})} = 1, \quad (21)$$

so $\pi(\sigma^2 | \mathbf{y}, \boldsymbol{\theta})$, as defined by equation (19), is a probability density. As a matter of fact, a simple calculation shows it is the density of the inverse-Gamma probability distribution with shape parameter $n/2$ and rate parameter $\frac{1}{2} \mathbf{y}^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{y}$.

From these preliminary calculations, we can derive the predictive distribution of unknown values \mathbf{y}_0 at unobserved points \mathbf{x}_0 knowing observed values \mathbf{y} :

$$\begin{aligned} P(\mathbf{y}_0 | \mathbf{y}) &= \int \int L(\mathbf{y}_0 | \mathbf{y}, \sigma^2, \boldsymbol{\theta}) \pi(\sigma^2 | \mathbf{y}, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} | \mathbf{y}) d\sigma^2 d\boldsymbol{\theta} \\ &= \int \int L(\mathbf{y}_0 | \mathbf{y}, \sigma^2, \boldsymbol{\theta}) \frac{L(\mathbf{y} | \sigma^2, \boldsymbol{\theta}) \pi(\sigma^2 | \boldsymbol{\theta})}{L^1(\mathbf{y} | \boldsymbol{\theta})} \pi(\boldsymbol{\theta} | \mathbf{y}) d\sigma^2 d\boldsymbol{\theta} \\ &= \int \left(\int L(\mathbf{y}_0, \mathbf{y} | \sigma^2, \boldsymbol{\theta}) \pi(\sigma^2) d\sigma^2 \right) L^1(\mathbf{y} | \boldsymbol{\theta})^{-1} \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}. \end{aligned} \quad (22)$$

The value up to $\pi(\sigma^2)$'s multiplicative constant of integral $\int L(\mathbf{y}_0, \mathbf{y} | \sigma^2, \boldsymbol{\theta}) \pi(\sigma^2) d\sigma^2$ is given by adapting equation (2). Denoting n_0 the length of \mathbf{y}_0 , $\boldsymbol{\Sigma}_{\theta,0,0}$ the associated correlation matrix, $\boldsymbol{\Sigma}_{\theta,0,\cdot}$ the correlation matrix between \mathbf{y}_0 and \mathbf{y} , and $\boldsymbol{\Sigma}_{\theta,\cdot,0}$ its transpose the correlation matrix between \mathbf{y} and \mathbf{y}_0 ,

$$\begin{aligned} L^{1,0}(\mathbf{y}_0, \mathbf{y} | \boldsymbol{\theta}) &:= \int L(\mathbf{y}_0, \mathbf{y} | \sigma^2, \boldsymbol{\theta}) \pi(\sigma^2) d\sigma^2 \\ &\propto \frac{\Gamma\left(\frac{n+n_0}{2}\right)}{\pi^{\frac{n+n_0}{2}}} \left| \begin{pmatrix} \boldsymbol{\Sigma}_{\theta,0,0} & \boldsymbol{\Sigma}_{\theta,0,\cdot} \\ \boldsymbol{\Sigma}_{\theta,\cdot,0} & \boldsymbol{\Sigma}_\theta \end{pmatrix} \right|^{-\frac{1}{2}} \left(\begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y} \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{\Sigma}_{\theta,0,0} & \boldsymbol{\Sigma}_{\theta,0,\cdot} \\ \boldsymbol{\Sigma}_{\theta,\cdot,0} & \boldsymbol{\Sigma}_\theta \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y} \end{pmatrix} \right)^{-\frac{n+n_0}{2}}. \end{aligned}$$

Finally,

$$P(\mathbf{y}_0 | \mathbf{y}) = \int \frac{L^{1,0}(\mathbf{y}_0, \mathbf{y} | \boldsymbol{\theta})}{L^1(\mathbf{y} | \boldsymbol{\theta})} \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}. \quad (23)$$

We can recognize $\frac{L^{1,0}(\mathbf{y}_0, \mathbf{y} | \boldsymbol{\theta})}{L^1(\mathbf{y} | \boldsymbol{\theta})}$ as the density of a multivariate Student t-distribution [Santner et al., 2003].

Proposition 6. $\frac{L^{1,0}(\mathbf{y}_0, \mathbf{y} | \boldsymbol{\theta})}{L^1(\mathbf{y} | \boldsymbol{\theta})}$, seen as a function of \mathbf{y}_0 , is the pdf of the multivariate Student t-distribution with n degrees of freedom, location vector $\boldsymbol{\Sigma}_{\theta,0,\cdot} \boldsymbol{\Sigma}_\theta^{-1} \mathbf{y}$ and scale matrix $\frac{\mathbf{y}^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{y}}{n} (\boldsymbol{\Sigma}_{\theta,0,0} - \boldsymbol{\Sigma}_{\theta,0,\cdot} \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\Sigma}_{\theta,\cdot,0})$.

It is in most cases sufficient to only derive the predictive distribution $P(\mathbf{y}_0 | \mathbf{y})$ at one point at a time (that is, when \mathbf{y}_0 is the unknown value taken by the Gaussian Process at one single point : $n_0 = 1$). For this specific purpose we derive the following corollary.

Corollary 7. Under the previously described model, assuming both \mathbf{y} and $\boldsymbol{\theta}$ to be known, if \mathbf{y}_0 is a lone unknown value of the Gaussian Process, then $z_0 := \sqrt{\frac{n}{\mathbf{y}^\top \boldsymbol{\Sigma}_\theta^{-1} \mathbf{y}}} \frac{\mathbf{y}_0 - \boldsymbol{\Sigma}_{\theta,0,\cdot} \boldsymbol{\Sigma}_\theta^{-1} \mathbf{y}}{\sqrt{1 - \boldsymbol{\Sigma}_{\theta,0,\cdot} \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\Sigma}_{\theta,\cdot,0}}}$ is a random variable following the Student t-distribution with n degrees of freedom.

If n exceeds 30, it is usually accepted that the Student t-distribution with n degrees of freedom can be approximated by the standard Normal distribution. As this threshold should be exceeded in practical cases, we would recommend performing all computations as though the multivariate Student t-distribution in proposition 6 were the multivariate Normal distribution. Naturally, the location vector would become the mean vector and the scale matrix the covariance matrix.

In practice, $P(\mathbf{y}_0 | \mathbf{y})$ can be accessed by sampling $\boldsymbol{\theta}$ according to posterior density $\pi(\boldsymbol{\theta} | \mathbf{y})$ as described in the previous subsection and then averaging all pdfs of the corresponding Student t-distributions or their Normal approximations. In the case of a one-dimensional value, the cumulative predictive distribution function $\int_{t \leq y_0} P(t | \mathbf{y}) dt$ can be obtained by averaging all cdfs of the corresponding Student t-distributions or their Normal approximations.

4.3 Simplifying Computations : the Maximum A Posteriori (MAP) estimator

The procedure described above turns out to have good predictive properties as will be seen in section 6 but is computationally intensive. One could prefer using the Maximum A Posteriori estimator (MAP) $\hat{\boldsymbol{\theta}}_{MAP}$ to estimate $\boldsymbol{\theta}$ instead of the Maximum Likelihood Estimator (MLE). Then $P(\mathbf{y}_0 | \mathbf{y})$ from equation (23) would be replaced by

$$\hat{P}_{MAP}(\mathbf{y}_0 | \mathbf{y}) = \frac{L^{1,0}(\mathbf{y}_0, \mathbf{y} | \hat{\boldsymbol{\theta}}_{MAP})}{L^1(\mathbf{y} | \hat{\boldsymbol{\theta}}_{MAP})}. \quad (24)$$

Thanks to proposition 6, $\hat{P}_{MAP}(\mathbf{y}_0 | \mathbf{y})$ is a Student t-distribution. Note that because using the MLE estimator for σ^2 is equivalent to using integrated likelihood L^1 as defined by (2), the predictive density associated with the MLE $(\hat{\sigma}_{MLE}^2, \hat{\boldsymbol{\theta}}_{MLE})$ has the same expression (with $\hat{\boldsymbol{\theta}}_{MLE}$ replacing $\hat{\boldsymbol{\theta}}_{MAP}$) and is also a Student t-distribution.

Although we claim that the MAP estimator is more robust than the MLE estimator (see test cases in section 5), it has a significant drawback when compared to the approach which makes full use of the posterior distribution : one may artificially change the MAP by changing the parametrization of the correlation lengths. Indeed, Gu [2016] argues that some specific parametrizations give better results than others. Such a phenomenon cannot occur when making full use of our posterior distribution. This is because the posterior distribution associated to the Jeffreys-rule prior distribution does not depend on the parametrization, but its mode does. To be more precise, the posterior distribution associated to the Gibbs reference prior does depend on the parametrization, because only conditional posterior distributions are associated to a Jeffreys-rule prior. The Gibbs reference prior is however invariant by any reparametrization of the type $f(\boldsymbol{\theta}) = (f_1(\theta_1), \dots, f_r(\theta_r))$.

5 Comparisons between the MLE and MAP estimators

5.1 Methodology

In this section, we compare the MLE and MAP estimators for accuracy and robustness.

Our test cases are 3-dimensional Gaussian Processes with Matérn anisotropic geometric correlation kernels with smoothness 5/2. Their mean is the null function, which only leaves us with the matter of estimating their correlation length for each dimension.

We use uniform designs : our observation points are randomly generated according to the uniform distribution in a cube with side length 1.

In order to measure the performance of our estimators, we define a suitable distance between two vectors of correlation lengths. Then the error of an estimator is defined as its distance to the “true” vector of correlation lengths.

Let g be the function such that for any t in $(-1, 1)$, $g(t) = \operatorname{argtanh}(t)$ and $g(-1) = g(1) = 0$. We use the convention that, for any matrix \mathbf{M} with elements in $[0, 1]$, $g(\mathbf{M})$ is the matrix resulting from applying g to every element of \mathbf{M} .

Definition 8. *For a given design set, the distance between two vectors of correlation lengths $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ is $\|g(\boldsymbol{\Sigma}_{\boldsymbol{\theta}^1}) - g(\boldsymbol{\Sigma}_{\boldsymbol{\theta}^2})\|$, where $\|\cdot\|$ denotes the Frobenius norm.*

This distance involves applying the Fisher transformation [Hotelling, 1953] (that is, the inverse hyperbolic tangent function) to every (non-1) correlation coefficient in both associated correlation matrices. This is a variance-stabilizing transformation. For any random variables U and V following the normal distribution with mean 0 and variance 1, let us denote $-1 < \rho < 1$ the correlation coefficient between U and V . If (U_i, V_i) ($1 \leq i \leq N$) are independent copies of (U, V) , then $\hat{\rho} = \sum_{i=1}^N U_i V_i / n$ is a random variable and $\operatorname{argtanh}(\hat{\rho})$ follows the normal distribution with mean $\operatorname{argtanh}(\rho)$ and variance $1/(N-3)$. So $\operatorname{argtanh}(\hat{\rho})$ ’s variance does not depend on ρ , whereas $\hat{\rho}$ ’s does and goes to zero for $|\rho| \rightarrow 1$. Involving the Fisher transformation in the definition of the distance between two vectors of correlation lengths is therefore a way to assert that vectors of correlation lengths can be far apart even if they both lead to highly correlated observations.

This allows us to make sure errors made when estimating near-1 correlation coefficients are no less taken into account than errors made when estimating near-0 correlation coefficients.

Let us choose a “true” vector of correlation lengths (and also a variance parameter, but this parameter has no effect on either the MLE or the MAP). Then we need to :

1. Sample n points randomly according to the uniform distribution on the unit cube (in the following, $n = 30$).
2. Generate the observations of the Gaussian Process at the sampled points according to the selected “true” variance and correlation lengths.
3. Compute the MLE and the MAP of the vector of correlation lengths and their errors.
4. Repeat steps 1 to 3 $m - 1$ times (in the following, $m = 500$).

This method allows us to derive an approximate distribution of the errors of both estimators when both the realization of the Gaussian Process and the design set vary. Thus we get to test the robustness of both estimators vs the variability of both the Gaussian Process and the choice of design set.

5.2 Results

This subsection provides results obtained on 3-dimensional Gaussian Processes with null mean function and Matérn anisotropic geometric correlation kernels with smoothness $5/2$. The results are divided by “true” vectors of correlation lengths. In each case, we give in table 1 the empirical Root Mean Square Errors (RMSEs) of both MLE and MAP estimators as functions of varying instances of the Gaussian Process and uniform design sets on the unit cube.

Most of the “true” vectors of correlation lengths featured in Table 1 were selected in a way to showcase the behavior of both estimators in strongly anisotropic cases, but one (0.5 - 0.5 - 0.5) also showcases their behavior if the true kernel is actually isotropic. And the final one (0.8 - 1 - 0.9) is used to illustrate the performance in the case of a strongly correlated Gaussian Process : this case is fundamentally different from all others, because the Matérn anisotropic geometric family of correlation kernels is designed in such a way that the correlation length with greatest influence is the lowest. Informally speaking, it is enough that one correlation length should be near zero to make the whole process very uncorrelated, even should all other correlation lengths be very high.

In all studied cases, the MAP estimator was more robust than the MLE estimator : its RMSE (Root Mean Square Error) was between 9 and 15% lower, as showcased in Table 1.

Corr. lengths	MLE	MAP	– (%)
0.4 – 0.8 – 0.2	3.49	2.97	15
0.5 – 0.5 – 0.5	4.00	3.46	13
0.7 – 1.3 – 0.4	4.02	3.64	9
0.8 – 0.3 – 0.6	3.75	3.26	13
0.8 – 1.0 – 0.9	4.65	4.18	10

Table 1: RMSE of the MLE and MAP estimators for several “true” vectors of correlation lengths. The last column displays in percents the decrease of the RMSE of the MAP estimator with respect to the MLE.

To get a better sense of the distribution of the error when the design set and the realization of the Gaussian Process vary, we give in Figure 2 violin plots of the errors in the two most extreme case : very low correlation (0.4 – 0.8 – 0.2) and very high correlation (0.8 – 1.0 – 0.9)

6 Comparison of the predictive distributions associated with the estimators (MLE and MAP) and the full posterior distribution

6.1 Methodology

We use the same test cases as before. In this section, our goal is to assess the accuracy of prediction intervals associated with both estimators and with the full posterior distribution. We consider 95% intervals : the lower bound is the 2.5% quantile and the upper bound the 97.5% quantile of predictive distribution $\hat{P}_{MLE}(\mathbf{y}_0 | \mathbf{y})$, $\hat{P}_{MAP}(\mathbf{y}_0 | \mathbf{y})$ and $P(\mathbf{y}_0 | \mathbf{y})$. For the sake of comprehensiveness, we also consider predictive intervals associated with “true” predictive distribution $L(\mathbf{y}_0 | \mathbf{y}, \sigma^2, \boldsymbol{\theta})$, the predictive distribution we would use if we knew the correct values of parameters σ^2 and $\boldsymbol{\theta}$.

Let us choose a “true” vector of correlation lengths $\boldsymbol{\theta}$ (and also a variance parameter σ^2 , but this parameter has no effect on predictive accuracy). Then we do the following :

1. Sample n observation points randomly according to the uniform distribution on the unit cube (in the following, $n = 30$).

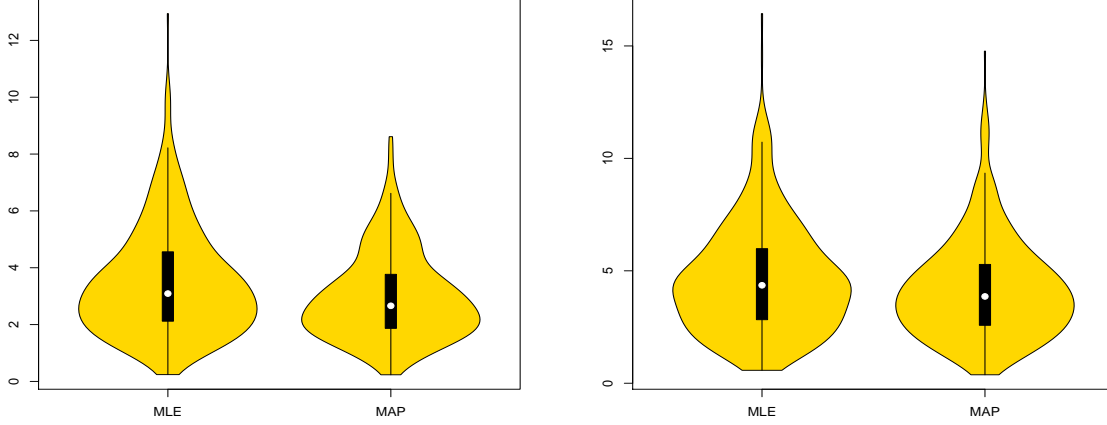


Figure 2: Violin plots of the error of the MLE and MAP estimators with respect to a design set following the uniform distribution and a Gaussian Process with correlation lengths $0.4 - 0.8 - 0.2$ (left) and $0.8 - 1.0 - 0.9$ (right).

2. Generate the observations of the Gaussian Process at the sampled points according to the selected “true” variance and correlation lengths.
3. Sample the vector of correlation lengths according to the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$ by Gibbs method.
4. Compute the MLE and the MAP of the vector of correlation lengths.
5. Sample n_0 test points randomly according to the uniform distribution on the unit cube (in the following, $n_0 = 100$).
6. At each point, determine the 95% prediction intervals derived from $L(\mathbf{y}_0 | \mathbf{y}, \sigma^2, \boldsymbol{\theta})$ (σ^2 and $\boldsymbol{\theta}$ being the “true” parameters), $\hat{P}_{MLE}(\mathbf{y}_0 | \mathbf{y})$, $\hat{P}_{MAP}(\mathbf{y}_0 | \mathbf{y})$ and $P(\mathbf{y}_0 | \mathbf{y})$.
7. Generate the values of the Gaussian Process at the newly sampled points (naturally, do this conditionally to the previously generated observations).
8. Count the number of points within the prediction intervals derived of each of the four predictive distributions. Divide the counts by n_0 : this yields four *coverages* corresponding to each type of predictive intervals. Also compute the *average length* of every type of prediction interval.
9. Repeat steps 1 to 8 $m - 1$ times (in the following, $m = 500$).

6.2 Results

There is no reason for individual coverages of 95% predictive intervals given by the predictive distribution to be equal to 95%. Recall that any coverage is given for a single realization of the Gaussian Process, and

that the values of this process at different points are correlated. If the predictive interval at some point fails to cover the true value at this point, it is likely that predictive intervals at neighboring points will also fail to cover the true values at those points, even though the nominal value is 95% everywhere. Conversely, if it actually covers the true value, then prediction intervals at neighboring points are more than 95% likely to cover their true values.

In short, prediction intervals give information that is only valid if understood to refer to what can be guessed on the sole basis of the observations made at the design points, which is why coverages for individual realizations of the Gaussian Process are not necessarily 95% *even if the predictive distribution is perfectly accurate* (*i.e.* based on the true values of σ^2 and θ).

However, *if the predictive distribution is perfectly accurate*, then the average of the coverages is the nominal value : 95%. It is thus interesting to compute the average of the coverages for all predictive distributions, whether they are based on the MLE or MAP estimator, or on the full posterior distribution (hereafter noted FPD). In the above described methodology, the average was taken over the realizations of the Gaussian Process with the chosen true parameters and over all design sets with n design points. The results below are obtained in this way.

The results given in Table 2 show that using the full posterior distribution (FPD) to derive the predictive distribution is the best possible choice from a frequentist point of view as the nominal value is nearly matched by the average coverage. Predictive intervals derived from the MAP estimator do not perform as well, and predictive intervals derived from the MLE perform even worse.

Corr. lengths	True	MLE	MAP	FPD
0.4 – 0.8 – 0.2	0.95	0.88	0.91	0.95
0.5 – 0.5 – 0.5	0.95	0.89	0.90	0.94
0.7 – 1.3 – 0.4	0.95	0.90	0.92	0.95
0.8 – 0.3 – 0.6	0.95	0.89	0.91	0.95
0.8 – 1.0 – 0.9	0.95	0.90	0.92	0.94

Table 2: Average with respect to randomly sampled design sets and realizations of the Gaussian Process (with variance parameter 1 and smoothness parameter 5/2) of the coverage of 95% Prediction Intervals across the sample space. “True” stands for the prediction based on the knowledge of the true variance parameter and the true vector of correlation lengths.

Let us now focus on the average (with respect to the uniform design sets and realizations of the Gaussian Process) of the mean (over the test set for a given realization of the Gaussian Process and a given uniform design set) length of prediction intervals. The results are given in Table 3, where the figures between parentheses give the increase or decrease (in percents) of the average mean length when compared to the average mean length of prediction intervals obtained using the true values of the parameters.

Predictive intervals derived from the full posterior distribution (FPD) are on average the largest, but not much larger than predictive intervals derived using the true parameters. In the tests we conducted, they seemed on average to be larger by about one fifth at worst. Predictive intervals derived from the MLE and

Corr. lengths	True	MLE	MAP	FPD
0.4 – 0.8 – 0.2	2.23	2.05 (-8)	2.13 (-4)	2.59 (+16)
0.5 – 0.5 – 0.5	1.69	1.55 (-8)	1.58 (-6)	1.84 (+9)
0.7 – 1.3 – 0.4	1.09	1.02 (-6)	1.07 (-2)	1.21 (+11)
0.8 – 0.3 – 0.6	1.63	1.51 (-7)	1.56 (-4)	1.82 (+12)
0.8 – 1.0 – 0.9	0.71	0.66 (-7)	0.69 (-3)	0.76 (+8)

Table 3: Average with respect to randomly sampled design sets and realizations of the Gaussian Process (with variance parameter 1 and smoothness parameter 5/2) of the coverage of 95% Prediction Intervals across the sample space. “True” stands for the prediction based on the knowledge of the true variance parameter and the true vector of correlation lengths.

MAP estimators are on average shorter than those derived from the true parameters. This can be interpreted as an under-estimation of the uncertainty of the prediction when fixing the vector of correlation lengths to the most likely value given the observations, and this can explain the low observed coverage in Table 2.

In figure 3, we give violin plots of coverage and mean length of Prediction Intervals in the two most extreme cases : correlation lengths 0.4 – 0.8 – 0.2 (very low correlation) and 0.8 – 1.0 – 0.9 (very high correlation). The results are similar and illustrate the fact that the FPD gives larger intervals in order to reach the derived coverage value.

6.3 A higher-dimensional case

In this subsection, we emulate using Simple Kriging the 10-dimensional Ackley function :

$$A(\mathbf{x}) = 20 + \exp(1) - 20 \exp \left(-0.2 \sqrt{\frac{1}{10} \sum_{i=1}^{10} x_i^2} \right) - \exp \left(\frac{1}{10} \sum_{i=1}^{10} \cos(2\pi x_i) \right). \quad (25)$$

The goal in this section is to emulate the Ackley function on the unit hypercube $[0, 1]^{10}$ using a design set with 100 observation points chosen through Latin Hypercube Sampling. The Simple Kriging model uses the null function as mean function and the Matérn anisotropic geometric covariance kernel family with smoothness parameter 5/2. The reference Gibbs posterior distribution is accessed through a sample of 1000 points. The conditional densities are sampled using the Metropolis algorithm with normal instrumental density with standard deviation 0.4 and a 20-step burn-in period, which as before leads to a 0.65 acceptance rate.

To evaluate the performance of prediction intervals, we follow steps 3, 4, 5, 6 and 8 of the method presented in this section (step 7 is skipped as the “values of the Gaussian process” are naturally the values of the Ackley function) with $n_0 = 1000$. The results are presented in Table 4. In this specific case, using the full posterior distribution does not seem to improve performance over the MAP estimate, but both perform significantly better than the MLE estimate.

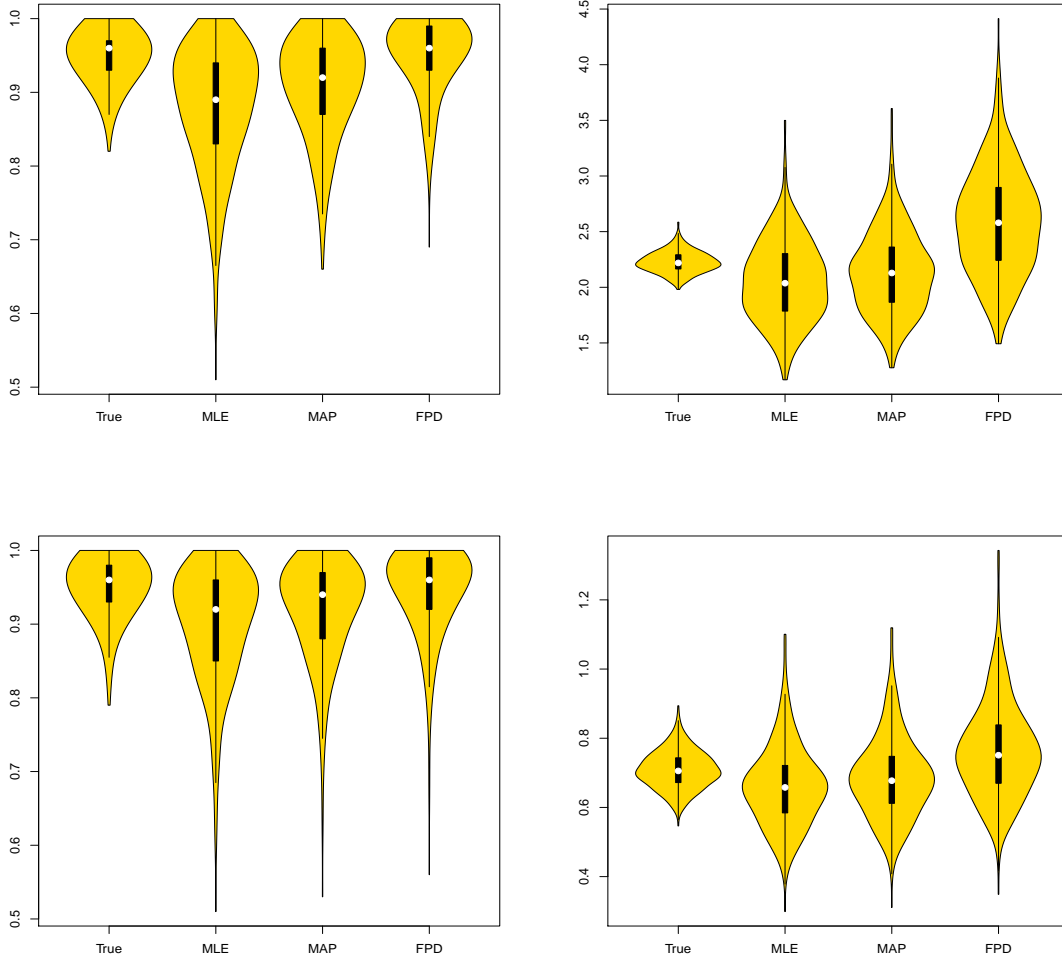


Figure 3: Violin plots of the coverage (left) and mean length (right) of Prediction Intervals with respect to a design set following the uniform distribution and a Gaussian Process with correlation lengths 0.4 – 0.8 – 0.2 (top) and 0.8 – 1.0 – 0.9 (bottom).

95 % Prediction intervals	MLE	MAP	FPD
Coverage	0.88	0.95	0.96
Mean length	0.31	0.36 (+16)	0.40 (+27)

Table 4: Coverage and mean length of 95 % prediction intervals when emulating the Ackley function on the unit hypercube using a Gaussian Process with null mean function and a Matérn anisotropic geometric covariance kernel with smoothness 5/2, unknown variance parameter and unknown vector of correlation lengths. The design set contains 100 points (Latin Hypercube Sampling). The figures in parentheses represent in percents the increase of the mean length when using the MAP/FPD with respect to the MLE.

7 Conclusion and Perspectives

We suggested a framework for deriving a new objective prior distribution, the Gibbs reference prior, on Simple Kriging parameters. Applying this framework to Matérn anisotropic kernels, we showed prediction to have good frequentist properties.

This framework is likely to apply to power exponential kernels [De Oliveira et al., 1997], with the possible exception of the squared exponential kernel. Indeed, notice the exponential kernel is also the Matérn kernel with smoothness parameter $\nu = 1/2$, while the squared exponential kernel is the limit of Matérn correlation kernels when $\nu \rightarrow \infty$. Rational quadratic kernels [Yaglom, 1987] may fit the framework.

The next step is to derive the Gibbs reference prior in the framework of Universal Kriging, which is the same model with a mean function that is unknown but assumed to be a linear combination of known functions f_1, \dots, f_p . The linear coefficients β_1, \dots, β_p are then considered parameters of the model. This extension is of practical relevance, because the mean function can rarely be considered known. It can probably be done in the same way Berger et al. [2001] extended the reference prior from the Simple Kriging to the Universal Kriging framework : they used the flat improper prior as joint prior on β_1, \dots, β_p conditional to σ^2 and $\boldsymbol{\theta}$ and used it to integrate β_1, \dots, β_p out of the likelihood function, and then proceeded to derive the reference prior on σ^2 and $\boldsymbol{\theta}$ with respect to the integrated likelihood.

A further extension would involve deriving an objective prior on the smoothness parameter ν . This extension appears out of reach for now, even though the reference prior algorithm could theoretically be further applied by means of the Jeffreys-rule prior on ν . Putting aside the analytical difficulties of the operation, one should take into account the relationship between correlation length $\boldsymbol{\theta}$ and smoothness ν . Unfortunately, asymptotic theory is not of much help in this regard, as Anderes [2010] shows that provided the spatial domain D is of dimension at least 5, then all parameters of the Matérn anisotropic geometric kernel are microergodic (Zhang [2004] shows this to be untrue for spatial domains of dimension 1, 2 or 3, but the non-microergodic parameters are σ^2 and $\boldsymbol{\theta}$, not ν). This means that the Gaussian measures on D corresponding to Gaussian Processes with two different smoothness parameters are orthogonal, which suggests that there exists a consistent estimator (the MLE possibly). Stein [1999] (section 6.6) considers the Fisher information on $\boldsymbol{\theta}$ and ν , and gives examples (with a one-dimensional sample space D) showing that the Fisher information on these parameters depends a lot on the design set. Fisher information relative to the smoothness parameter ν increases when design points are chosen to be close to one another (relative to the "true" correlation length $\boldsymbol{\theta}$), whereas Fisher information relative to correlation length $\boldsymbol{\theta}$ is maximized for design points that are farther apart. This, according to him, is coherent with the fact that $\boldsymbol{\theta}$ has greater influence on the low frequency behavior of the Matérn kernel while ν has greater influence on its high frequency behavior. This also suggests to us that the smoothness parameter ν , like the variance parameter σ^2 , can only be meaningfully estimated if the vector of correlation lengths $\boldsymbol{\theta}$ is known. Otherwise, the estimator could hardly tell which design points are close to each other, which intuitively seems a prerequisite to evaluating the smoothness of the process. If we wish to apply the reference prior algorithm to the case where ν is unknown, we should thus probably derive the reference prior on ν conditional to $\boldsymbol{\theta}$ to obtain in the end a prior of the type $\pi(\sigma^2 \mid \nu, \boldsymbol{\theta})\pi(\nu \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Appendices

A Assumption on Σ_θ 's asymptotic expansion

Berger et al. [2001] consider isotropic kernels (*i.e.* with a one-dimensional correlation length θ). For the sake of convenience, we adopt the same framework in this section. In their Lemma 2, Berger et al. [2001] suppose that the used correlation kernel and design set are such that, as $\theta \rightarrow \infty$, $\Sigma_\theta = \mathbf{1}\mathbf{1}^\top + \nu(\theta)\mathbf{D} + \mathbf{R}(\theta)$, where $\mathbf{1}$ is the vector with n entries all equal to 1, $\nu(\theta)$ is a real-valued function such that $\nu(\theta) \rightarrow 0$, \mathbf{D} is a fixed nonsingular matrix and $\mathbf{R}(\theta)$ is a matrix such that $\|\frac{1}{\nu(\theta)}\mathbf{R}(\theta)\| \rightarrow 0$.

What makes this assumption restrictive is the condition that \mathbf{D} should be nonsingular, because this is only true for rather irregular correlation kernels. For instance, as was noted by Paulo [2005], it is untrue for squared exponential correlation kernels.

For a given correlation kernel and design set $\{\mathbf{x}^{(i)}, i = 0, 1, \dots, n\}$, \mathbf{D} is typically a matrix proportional to the matrix with entries $\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^p$, where p depends on the smoothness of the correlation kernel but should in any case belong to the interval $(0, 2]$. Schoenberg [1937] gives the following result (Theorem 4 in the original paper) :

Theorem 9. *If $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ ($n \geq 1$) are distinct points of a Euclidean space $(E, \|\cdot\|)$, the quadratic form*

$$q(\xi) = \sum_{i,j=0}^n \left\| \mathbf{x}^{(i)} - \mathbf{x}^{(j)} \right\|^p \xi_i \xi_j \quad (26)$$

with $p \in (0, 2)$ is nonsingular and its canonical representation contains one positive and n negative squares.

This means that, if the correlation kernel is irregular enough to have $0 < p < 2$, the assumption that \mathbf{D} is nonsingular is reasonable.

Corollary 10. *If $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ ($n \geq 1$) are distinct points of a Euclidean space $(E, \|\cdot\|)$ the matrix with entries $\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^p$ with $p \in (0, 2)$ is nonsingular and has one positive eigenvalue and n negative eigenvalues.*

Things are different when the correlation kernel is regular enough to have $p = 2$, however. Gower [1985]'s Theorem 6 implies the following results :

Theorem 11. *If $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ ($n \geq 1$) are distinct points of a Euclidean space $(E, \|\cdot\|)$ and d is the dimension of E_d , the smallest Euclidean subspace containing them, the matrix with entries $\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2$ has rank :*

- $d + 1$ (one positive eigenvalue, d negative eigenvalues, any other eigenvalue null) if $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ lie on the surface of a hypersphere of E_d ;
- $d + 2$ (one positive eigenvalue, $d + 1$ negative eigenvalues, any other eigenvalue null) otherwise.

Corollary 12. *If $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ ($n \geq 1$) are distinct points of a Euclidean space $(E, \|\cdot\|)$, the matrix with entries $\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2$ has rank lower than the dimension of the smallest Euclidean space containing them plus 3.*

For all practical purposes, the number of design points is much greater than the dimension of the spatial domain D , so the matrix \mathbf{D} is singular when $p = 2$.

Let us review the values of p corresponding to a few commonly used correlation kernels. Matérn correlation kernels [Matérn, 1986] [Handcock and Stein, 1993] with smoothness parameter ν have $p = 2 \min(1, \nu)$, thus for $0 < \nu < 1$, $0 < p < 2$ but for $\nu \geq 1$, $p = 2$. Spherical correlation kernels [Wackernagel, 1995] have $p = 1$. Power exponential kernels [De Oliveira et al., 1997] have p equal to their power. This means that all power exponential kernels except the squared exponential correlation kernel have $0 < p < 2$. In particular, the exponential kernel (which is also the Matérn kernel with smoothness $\nu = 1/2$) has $p = 1$, but the squared exponential kernel has $p = 2$. Rational quadratic kernels [Yaglom, 1987] have $p = 2$.

Note finally that, if some tensorized anisotropic correlation kernel is used, $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ can be written $\mathbf{\Sigma}_{\boldsymbol{\theta}} = \mathbf{\Sigma}_{\theta_1} \circ \dots \circ \mathbf{\Sigma}_{\theta_r}$ (\circ being the Hadamard product on matrices), and the same discussion could be held for every $\mathbf{\Sigma}_{\theta_k}$ ($1 \leq k \leq r$) when $\theta_k \rightarrow \infty$, provided the design set has coordinate-distinct points in the sense defined at the beginning of Section 3.

B Asymptotic behavior of $\boldsymbol{\theta} \mapsto L^1(\mathbf{y} \mid \boldsymbol{\theta})$

How does $L^1(\mathbf{y} \mid \boldsymbol{\theta})$ behave when correlation length $\boldsymbol{\theta}$ varies ? It is impossible to answer in a general framework, but $(\mathbf{y}^\top \mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{y})^{-\frac{n}{2}}$ increases when the correlation matrix determined by $\boldsymbol{\theta}$ evolves in such a way that its eigenvectors with lowest eigenvalues “become orthogonal” to \mathbf{y} (that is to say the scalar product of the unit eigenvectors with lowest eigenvalues and \mathbf{y} decreases). And $|\mathbf{\Sigma}_{\boldsymbol{\theta}}|^{-\frac{1}{2}}$ increases when eigenvalues of correlation matrix $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ decrease ; in other words, when the associated Gaussian distribution tends to concentrate around a hyperplane.

Is it possible for $L^1(\mathbf{y} \mid \boldsymbol{\theta})$ to increase to infinity as $\boldsymbol{\theta}$ varies ? $(\mathbf{y}^\top \mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{y})^{-\frac{n}{2}}$ is bounded due to $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ being a correlation matrix . $|\mathbf{\Sigma}_{\boldsymbol{\theta}}|^{-\frac{1}{2}}$ should therefore increase to infinity when $\boldsymbol{\theta}$ goes to infinity, which would imply for $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ ’s lowest eigenvalue to go to zero and for the Gaussian distribution to move ever closer to degeneracy. However, as $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ gets closer to degeneracy, $(\mathbf{y}^\top \mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{y})^{-\frac{n}{2}}$ is likely to crash to zero about as fast as $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ ’s lowest value taken at the power $n/2$, which is at least as fast the rate of $|\mathbf{\Sigma}_{\boldsymbol{\theta}}|^{-\frac{1}{2}}$ ’s ascent to infinity. To prevent that, some unit eigenvector of $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ with lowest eigenvalue should get ever closer to being orthogonal to \mathbf{y} , that is, its scalar product with \mathbf{y} should go to zero.

Thus for $L^1(\mathbf{y} \mid \boldsymbol{\theta})$ not to be bounded would mean there exists a degenerate Gaussian distribution spanning a hyperplane of \mathbb{R}^n that accounts for \mathbf{y} . If that were truly the case, then this distribution would be a better candidate than any nondegenerate Gaussian distribution and the Kriging model would be inappropriate.

For this reason, the assumption that $L^1(\mathbf{y} \mid \boldsymbol{\theta})$ remains bounded as $\boldsymbol{\theta}$ varies seems reasonable.

C Proofs of section 2

Proof of proposition 1. As $\mathbf{M}_{\boldsymbol{\theta}}^{\Sigma}$ is a symmetric matrix, the spectral theorem guarantees the existence of a diagonal matrix $\mathbf{\Lambda}_{\boldsymbol{\theta}}^{\Sigma}$ and an orthogonal matrix $\mathbf{O}_{\boldsymbol{\theta}}^{\Sigma}$ such that $\mathbf{M}_{\boldsymbol{\theta}}^{\Sigma} = (\mathbf{O}_{\boldsymbol{\theta}}^{\Sigma})^\top \mathbf{\Lambda}_{\boldsymbol{\theta}}^{\Sigma} (\mathbf{O}_{\boldsymbol{\theta}}^{\Sigma})$, with the di-

agonal coefficients of $\mathbf{\Lambda}_\theta^\Sigma$ being the eigenvalues of \mathbf{M}_θ^Σ . Setting $\mathbf{U}_0 := (\mathbf{O}_\theta^\Sigma)\mathbf{U}$, we can now compute $\text{Var}[\mathbf{U}_0^\top \mathbf{\Lambda}_\theta^\Sigma \mathbf{U}_0] = \text{Var}[\mathbf{U}^\top \mathbf{M}_\theta^\Sigma \mathbf{U}]$, \mathbf{U}_0 following the uniform distribution on S^{n-1} .

Let $(\lambda_i)_{1 \leq i \leq n}$ be the eigenvalues of \mathbf{M}_θ^Σ . We can write $\text{Var}[\mathbf{U}_0^\top \mathbf{\Lambda}_\theta^\Sigma \mathbf{U}_0] = \text{Var}[\sum_{1 \leq i \leq n} \lambda_i X_i]$, where X_i ($1 \leq i \leq n$) are nonnegative identically distributed random variables such that $\sum_{1 \leq i \leq n} X_i = 1$.

$$\text{Var}\left[\sum_{i=1}^n \lambda_i X_i\right] = \text{Var}[X_1] \sum_{i=1}^n \lambda_i^2 + 2 \text{Cov}(X_1, X_2) \sum_{1 \leq i < j \leq n} \lambda_i \lambda_j. \quad (27)$$

Obviously, $\mathbb{E}[X_1] = \frac{1}{n}$ and thus $\text{Cov}(X_1, X_2) = -1/(n-1) \text{Var}[X_1]$.

$$\begin{aligned} \frac{\text{Var}[\sum_{i=1}^n \lambda_i X_i]}{\text{Var}[X_1]} &= \sum_{i=1}^n \lambda_i^2 - \frac{1}{n-1} \sum_{i=1}^n \lambda_i \sum_{j \neq i} \lambda_j \\ &= \left(1 + \frac{1}{n-1}\right) \left(\text{Tr}[(\mathbf{M}_\theta^\Sigma)^2] - \frac{1}{n} \text{Tr}[\mathbf{M}_\theta^\Sigma]^2\right). \end{aligned} \quad (28)$$

□

Proof of Theorem 3. If such a proper distribution exists, it is obviously unique, as it can be characterized using its conditional marginal distributions. It is thus enough to show its existence, which will be done using a recursive argument. The recurrence assumption $H(R)$ is defined thus :

$H(R)$: If $r \geq R$, then whatever the value of the vector $(\theta_j)_{R+1 \leq j \leq r}$, one can define a probability density $\pi((\theta_i)_{1 \leq j \leq R} = \cdot \mid (\theta_j)_{R+1 \leq j \leq r})$ (with the convention that if $r = R$, then $\pi((\theta_i)_{1 \leq j \leq R} = \cdot \mid (\theta_j)_{R+1 \leq j \leq r}) = \pi((\theta_i)_{1 \leq j \leq R} = \cdot)$) such that for every integer i verifying $1 \leq i \leq R$, whatever the value of the vector $(\theta_j)_{1 \leq j \leq R}$, the conditional density $\pi(\theta_i = \cdot \mid \theta_j \forall j \neq i)$ is proportional to $d(\theta_i = \cdot \mid \theta_j \forall j \neq i)$.

Let us first show that $H(R) \implies H(R+1)$. If $r \geq R+1$ and $H(R)$ holds, first set $\theta_{R+2}, \dots, \theta_r$ (if $R+1 = r$, this is unnecessary with our convention). Then define $f_{R+1}(\theta_{R+1} = \cdot \mid \theta_{R+2}, \dots, \theta_r)$ such that for every $\tau \in \mathbb{R}_+$,

$$\frac{1}{f_{R+1}(\theta_{R+1} = \tau \mid \theta_{R+2}, \dots, \theta_r)} = \int_0^\infty \dots \int_0^\infty \frac{\pi(\theta_1 = t_1, \dots, \theta_R = t_R \mid \theta_{R+1} = \tau, \theta_{R+2}, \dots, \theta_r)}{d(\theta_{R+1} = \tau \mid \theta_1 = t_1, \dots, \theta_R = t_R, \theta_{R+2}, \dots, \theta_r)} dt_1 \dots dt_R. \quad (29)$$

The numerator in the right member of the above equation is well defined thanks to $H(R)$. Thus, because $\pi(\theta_1, \dots, \theta_R \mid \theta_{R+1}, \theta_{R+2}, \dots, \theta_r)$ is a probability density on $(\mathbb{R}_+)^R$, Lemma 2 yields

$$f_{R+1}(\theta_{R+1} = \tau \mid \theta_{R+2}, \dots, \theta_r) \leq M_{R+1} \min\left(1, \frac{1}{\tau^b}\right). \quad (30)$$

As $b > 1$, this shows that $f_{R+1}(\theta_{R+1} = \cdot \mid \theta_{R+2}, \dots, \theta_r)$ is the density with respect to Lebesgue measure of a finite measure on \mathbb{R}_+^* . It is proportional to a probability density, which we will note $f_{R+1}^N(\theta_{R+1} = \cdot \mid \theta_{R+2}, \dots, \theta_r)$. We can then define for every value of $(\theta_i)_{1 \leq i \leq R+1} \in (\mathbb{R}_+)^{R+1}$

$$g_{R+1}^N(\theta_1, \dots, \theta_R, \theta_{R+1} \mid \theta_{R+2}, \dots, \theta_r) := \pi(\theta_1, \dots, \theta_R \mid \theta_{R+1}, \theta_{R+2}, \dots, \theta_r) f_{R+1}^N(\theta_{R+1} \mid \theta_{R+2}, \dots, \theta_r). \quad (31)$$

Clearly, $g_{R+1}^N(\theta_1 = \cdot, \dots, \theta_R = \cdot, \theta_{R+1} = \cdot | \theta_{R+2}, \dots, \theta_r)$ is a probability density with respect to Lebesgue measure on $(\mathbb{R}_+^*)^{R+1}$. $H(R)$ and equation (31) show that for every integer $1 \leq i \leq R$, whatever $\theta_j > 0$ ($1 \leq j \leq R$ and $j \neq i$), $g_{R+1}^N(\theta_i = \cdot | \theta_j \forall j \neq i)$ is proportional to $d(\theta_i = \cdot | \theta_j \forall j \neq i)$. Moreover, equation (29) shows that whatever $\theta_j > 0$ ($1 \leq j \leq R$), $g_{R+1}^N(\theta_{R+1} = \cdot | \theta_j \forall j \neq R+1)$ is proportional to $d(\theta_{R+1} = \cdot | \theta_j \forall j \neq R+1)$.

Thus, setting $\pi(\theta_1 = \cdot, \dots, \theta_R = \cdot, \theta_{R+1} = \cdot | \theta_{R+2}, \dots, \theta_r) = g_{R+1}^N(\theta_1 = \cdot, \dots, \theta_R = \cdot, \theta_{R+1} = \cdot | \theta_{R+2}, \dots, \theta_r)$ we have $H(R+1)$.

The proof of $H(2)$ is actually the same with the following definition : $\pi(\theta_1 = \cdot | \theta_j \forall 2 \leq j \leq r)$ is the probability density proportional to $d(\theta_1 = \cdot | \theta_j \forall 2 \leq j \leq r)$. \square

Proof of Theorem 4. This result can be derived from equation (30) in the proof of Proposition 3. \square

D Proofs of section 3

This section is devoted to proving Proposition 5. To do this, we need to prove all 6 majorations in assumption 1. Subsection D.1 is devoted to proving the first 4 majorations, culminating in Proposition 15. Subsection D.2, meanwhile, is dedicated to proving majorations 5 and 6, although the main difficulty lies in proving Proposition 20 which gives majoration 6, majoration 5 being trivial.

D.1 Proof of majorations 1-4

Lemma 13. *The partial derivative with respect to θ_i of the Matérn tensorized kernel of variance σ^2 , smoothness ν and correlation length vector $\boldsymbol{\theta}$ is :*

$$\frac{\partial}{\partial \theta_i} \left(\sigma^2 K_{r,\nu}^{tens} \left(\frac{\mathbf{x}}{\boldsymbol{\theta}} \right) \right) = \frac{\sigma^2 (2\sqrt{\nu})^2 |x_i|^2}{\Gamma(\nu) 2^{\nu-1} \theta_i^3} \left(2\sqrt{\nu} \frac{|x_i|}{\theta_i} \right)^{\nu-1} B_{\nu-1} \left(2\sqrt{\nu} \frac{|x_i|}{\theta_i} \right) \prod_{j \neq i} K_{1,\nu} \left(\frac{|x_j|}{\theta_j} \right) \quad (32)$$

This can be rewritten as :

$$\frac{\partial}{\partial \theta_i} \left(\sigma^2 K_{r,\nu}^{tens} \left(\frac{\mathbf{x}}{\boldsymbol{\theta}} \right) \right) = \begin{cases} \sigma^2 \frac{2\nu}{\nu-1} \frac{|x_i|^2}{\theta_i^3} K_{\nu-1} \left(\frac{|x_i|}{\theta_i} \right) & \prod_{j \neq i} K_{1,\nu} \left(\frac{|x_j|}{\theta_j} \right) \text{ if } \nu > 1 \\ \sigma^2 4 \frac{|x_i|^2}{\theta_i^3} B_0 \left(2 \frac{|x_i|}{\theta_i} \right) & \prod_{j \neq i} K_{1,\nu} \left(\frac{|x_j|}{\theta_j} \right) \text{ if } \nu = 1 \\ \sigma^2 2\nu^\nu \frac{\Gamma(1-\nu)}{\Gamma(\nu)} \frac{|x_i|^{2\nu}}{\theta_i^{1+2\nu}} K_{1-\nu} \left(\frac{|x_i|}{\theta_i} \right) & \prod_{j \neq i} K_{1,\nu} \left(\frac{|x_j|}{\theta_j} \right) \text{ if } \nu < 1 \end{cases} \quad (33)$$

Proof. The first assertion is a simple matter of differentiating equation (16). In the following calculation,

the fourth line is given by formula 9.6.28 (page 376) in Abramowitz and Stegun [1964].

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} \left(\sigma^2 K_{r,\nu}^{tens} \left(\frac{\mathbf{x}}{\boldsymbol{\theta}} \right) \right) &= \sigma^2 \frac{\partial}{\partial \theta_i} \left(K_{1,\nu} \left(\frac{x_i}{\theta_i} \right) \right) \prod_{j \neq i} K_{1,\nu} \left(\frac{|x_j|}{\theta_j} \right) \\
&= \sigma^2 \frac{-x_i}{\theta_i^2} \left(K'_{1,\nu} \left(\frac{x_i}{\theta_i} \right) \right) \prod_{j \neq i} K_{1,\nu} \left(\frac{|x_j|}{\theta_j} \right) \\
&= \sigma^2 \frac{-x_i}{\theta_i^2} \left(\frac{2\sqrt{\nu}}{\Gamma(\nu)2^{\nu-1}} \frac{d}{dy} \Big|_{y=2\sqrt{\nu} \frac{x_i}{\theta_i}} [y^\nu B_\nu(y)] \right) \prod_{j \neq i} K_{1,\nu} \left(\frac{|x_j|}{\theta_j} \right) \\
&= \sigma^2 \frac{-x_i}{\theta_i^2} \left(\frac{2\sqrt{\nu}}{\Gamma(\nu)2^{\nu-1}} [-y \cdot y^{\nu-1} B_{\nu-1}(y)]_{y=2\sqrt{\nu} \frac{x_i}{\theta_i}} \right) \prod_{j \neq i} K_{1,\nu} \left(\frac{|x_j|}{\theta_j} \right)
\end{aligned} \tag{34}$$

From there, equation (32) follows immediately. Rewriting it in the form given in (33) only requires us to recall $\Gamma(\nu) = (\nu-1)\Gamma(\nu-1)$ (case $\nu > 1$), $\Gamma(1) = 1$ (case $\nu = 1$) and $B_{\nu-1} = B_{1-\nu}$ (case $\nu < 1$). \square

Lemma 14. *The partial derivative with respect to θ_i of the Matérn geometric anisotropic kernel of variance σ^2 , smoothness ν and correlation length vector $\boldsymbol{\theta}$ is :*

$$\frac{\partial}{\partial \theta_i} \left(\sigma^2 K_{r,\nu} \left(\frac{\mathbf{x}}{\boldsymbol{\theta}} \right) \right) = \frac{\sigma^2 (2\sqrt{\nu})^2 |x_i|^2}{\Gamma(\nu)2^{\nu-1} \theta_i^3} \left(2\sqrt{\nu} \left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\| \right)^{\nu-1} B_{\nu-1} \left(2\sqrt{\nu} \left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\| \right) \tag{35}$$

This can be rewritten as :

$$\frac{\partial}{\partial \theta_i} \left(\sigma^2 K_{r,\nu} \left(\frac{\mathbf{x}}{\boldsymbol{\theta}} \right) \right) = \begin{cases} \sigma^2 \frac{2\nu}{\nu-1} \frac{|x_i|^2}{\theta_i^3} K_{\nu-1} \left(\left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\| \right) & \text{if } \nu > 1 \\ \sigma^2 4 \frac{|x_i|^2}{\theta_i^3} B_0 \left(2 \left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\| \right) & \text{if } \nu = 1 \\ \sigma^2 2\nu \frac{\Gamma(1-\nu)}{\Gamma(\nu)} \frac{1}{\theta_i} \left(\frac{|x_i|/\theta_i}{\left\| \mathbf{x}/\boldsymbol{\theta} \right\|^{1-\nu}} \right)^2 K_{1-\nu} \left(\left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\| \right) & \text{if } \nu < 1 \end{cases} \tag{36}$$

Proof. The first assertion is a simple matter of differentiating equation (14). In the following calculation, the fourth line is given by formula 9.6.28 (page 376) in Abramowitz and Stegun [1964].

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} \left(\sigma^2 K_{r,\nu} \left(\frac{\mathbf{x}}{\boldsymbol{\theta}} \right) \right) &= \sigma^2 \frac{\partial}{\partial \theta_i} \left(K_{1,\nu} \left(\left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\| \right) \right) \\
&= \sigma^2 \frac{-x_i^2}{\theta_i^3} \left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\|^{-1} K'_{1,\nu} \left(\left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\| \right) \\
&= \sigma^2 \frac{-x_i^2}{\theta_i^3} \left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\|^{-1} \left(\frac{2\sqrt{\nu}}{\Gamma(\nu)2^{\nu-1}} \frac{d}{dy} \Big|_{y=2\sqrt{\nu} \left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\|} [y^\nu B_\nu(y)] \right) \\
&= \sigma^2 \frac{-x_i^2}{\theta_i^3} \left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\|^{-1} \left(\frac{2\sqrt{\nu}}{\Gamma(\nu)2^{\nu-1}} [-y \cdot y^{\nu-1} B_{\nu-1}(y)]_{y=2\sqrt{\nu} \left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\|} \right)
\end{aligned} \tag{37}$$

From there, equation (35) follows immediately. Rewriting it in the form given in (36) only requires us to recall $\Gamma(\nu) = (\nu-1)\Gamma(\nu-1)$ (case $\nu > 1$), $\Gamma(1) = 1$ (case $\nu = 1$) and $B_{\nu-1} = B_{1-\nu}$ (case $\nu < 1$). \square

Proposition 15. *Both tensorized and geometric anisotropic Matérn kernels verify majorations 1-4 of assumption 1 with any $a \geq 0$ and with*

$$\begin{aligned}
b &\leq 3 & \text{if } \nu > 1 \\
b &< 3 & \text{if } \nu = 1 \\
b &\leq 1 + 2\nu & \text{if } \nu < 1
\end{aligned}$$

Proof. As we are dealing with correlation matrices and kernels, we set the variance parameter σ^2 to be 1. It this proof, μ is a generic (positive) value of the smoothness parameter of a Matérn correlation kernel.

For the sake of concision, we only fully prove the assertion in the case where $\nu > 1$, but we outline the ideas of the proofs for the cases where $\nu = 1$ and where $\nu < 1$.

The general method of the proof is to make sure every single element of the derivative of the correlation matrix with respect to θ_i (for any $1 \leq i \leq r$) is bounded. This can be done thanks to the two previous lemmas, which give us closed forms for the elements of this derivative with \mathbf{x} being the generic difference between the coordinates of two points of the design set (which is assumed to have coordinate-distinct points).

Let us start with the case where a tensorized Matérn correlation kernel with smoothness parameter $\nu > 1$ is used.

- Majoration 1 holds because :

- whatever $\mu > 0$, K_μ is a bounded function ;
- whatever $\mu > 0$, $c \geq 0$ and $1 \leq k \leq r$, if $x_k \neq 0$, then $\theta_k^{-c} K_\mu \left(\frac{|x_k|}{\theta_k} \right) \rightarrow 0$ when $\theta_k \rightarrow 0$;

- Majoration 2 holds because :

- of the arguments justifying majoration 1 ;
- $\frac{\theta_i^b}{\theta_i^3} = O(1)$ when $\theta_i \rightarrow \infty$, provided $b \leq 3$.

- Majoration 3 holds because :

- of the arguments justifying majoration 1 ;
- the design set has coordinate-distinct points, which ensures that if some $1 \leq k \leq r$ verifies $x_k = 0$, then they all do.

- Majoration 4 holds because of the arguments justifying majorations 1, 2 and 3.

This proof is also valid in the case where a geometric anisotropic Matérn correlation kernel with smoothness parameter $\nu > 1$ is used, once we have observed that $K_\mu \left(\left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\| \right) \leq \min_{1 \leq j \leq r} K_\mu \left(\frac{|x_j|}{\theta_j} \right)$ since K_μ is a decreasing function on \mathbb{R}_+^* and $\left\| \frac{\mathbf{x}}{\boldsymbol{\theta}} \right\| \geq \max_{1 \leq j \leq r} \left(\frac{|x_j|}{\theta_j} \right)$.

Case $\nu = 1$: For tensorized Matérn kernels, all four majorations can be derived from the following facts :

- $B_0(t) \sim -\log(t)$ when $t \rightarrow 0$ according to Abramowitz and Stegun [1964] (9.6.8 page 378);
- $B_0(t) = o\left(\frac{1}{t^n}\right)$ when $t \rightarrow \infty$ for any nonnegative real number n according to Abramowitz and Stegun [1964] (9.7.2 page 378);
- the design set has coordinate-distinct points, which ensures that if some $1 \leq k \leq r$ verifies $x_k = 0$, then they all do.

For geometric anisotropic kernels, we also need to observe that $B_0(2\|\frac{\mathbf{x}}{\boldsymbol{\theta}}\|) \leq \min_{1 \leq j \leq r} B_0\left(2\frac{|x_j|}{\theta_j}\right)$ since B_0 is a decreasing function on \mathbb{R}_+^* and $\|\frac{\mathbf{x}}{\boldsymbol{\theta}}\| \geq \max_{1 \leq j \leq r} \left(\frac{|x_j|}{\theta_j}\right)$.

Case $\nu < 1$: For tensorized Matérn kernels, the four majorations can be derived in much the same way as in the case where $\nu > 1$. The same is true for anisometric geometric Matérn kernels after observing that, since $\|\frac{\mathbf{x}}{\boldsymbol{\theta}}\| \geq \max_{1 \leq j \leq r} \left(\frac{|x_j|}{\theta_j}\right)$, we have $\left(\frac{|x_i|/\theta_i}{\|\mathbf{x}/\boldsymbol{\theta}\|^{1-\nu}}\right)^2 \leq \frac{|x_i|^{2\nu}}{\theta_i^{2\nu}}$. \square

D.2 Proof of majorations 5-6

It is obvious that any design set with coordinate-distinct points will guarantee majoration 5 of assumption 1 for Matérn tensorized and anisometric geometric cases. Only majoration 6 remains to prove.

Lemma 16. *There exists a correlation kernel $\tilde{K}_{r,\nu}$ such that, whatever the design set $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$, whatever $\boldsymbol{\theta} \in (\mathbb{R}_+^*)^r$ and $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n) \in \mathbb{R}^n$,*

$$\sum_{j,k=1}^n \xi_j \xi_k K_{r,\nu} \left(\frac{\mathbf{x}^{(j)} - \mathbf{x}^{(k)}}{\boldsymbol{\theta}} \right) \geq 2^{-\frac{r}{2}-\nu} M_r(\nu) f_{r,\nu}(\theta_r) \sum_{j,k=1}^n \xi_j \xi_k \tilde{K}_{r,\nu} \left(\frac{\theta_r}{\boldsymbol{\theta}} (\mathbf{x}^{(j)} - \mathbf{x}^{(k)}) \right) \quad (38)$$

where $f_{r,\nu}(\theta) = (2\sqrt{\nu})^{-r-2\nu}\theta^r$ if $\theta \leq 2\sqrt{\nu}$ and $f_{r,\nu}(\theta) = \theta^{-2\nu}$ if $\theta \geq 2\sqrt{\nu}$.

Proof. Whatever $\mathbf{x}, \mathbf{y} \in \mathbb{R}^r$, $K_{r,\nu}(\mathbf{x} - \mathbf{y}) = \int_{\mathbb{R}^r} \widehat{K}_{r,\nu}(\boldsymbol{\omega}) e^{i\langle \boldsymbol{\omega} | \mathbf{x} - \mathbf{y} \rangle} d\boldsymbol{\omega}$.

$$\begin{aligned} \sum_{j,k=1}^n \xi_j \xi_k K_{r,\nu} \left(\frac{\mathbf{x}^{(j)} - \mathbf{x}^{(k)}}{\boldsymbol{\theta}} \right) &= \int_{\mathbb{R}^r} \widehat{K}_{r,\nu}(\boldsymbol{\omega}) \left| \sum_{j=1}^n \xi_j e^{i\langle \boldsymbol{\omega} | \frac{\mathbf{x}^{(j)}}{\boldsymbol{\theta}} \rangle} \right|^2 d\boldsymbol{\omega} \\ &= M_r(\nu) \underline{\theta}_r^r \int_{\mathbb{R}^r} (4\nu + \underline{\theta}_r^2 \|\mathbf{s}\|^2)^{-\frac{r}{2}-\nu} \left| \sum_{j=1}^n \xi_j e^{i\langle \frac{\theta_r}{\boldsymbol{\theta}} \mathbf{s} | \mathbf{x}^{(j)} \rangle} \right|^2 d\mathbf{s} \\ &\geq 2^{-\frac{r}{2}-\nu} M_r(\nu) f_{r,\nu}(\theta_r) \int_{\mathbb{R}^r \setminus B(0,1)} \|\mathbf{s}\|^{-r-2\nu} \left| \sum_{j=1}^n \xi_j e^{i\langle \frac{\theta_r}{\boldsymbol{\theta}} \mathbf{s} | \mathbf{x}^{(j)} \rangle} \right|^2 d\mathbf{s} \end{aligned} \quad (39)$$

Now, let $\tilde{K}_{r,\nu}$ be the function with Fourier transform $\widehat{\tilde{K}}_{r,\nu}(\boldsymbol{\omega}) = \mathbf{1}_{\{\|\boldsymbol{\omega}\| \geq 1\}} \|\boldsymbol{\omega}\|^{-r-2\nu}$. According to Bochner's theorem, $\tilde{K}_{r,\nu}$ is a correlation kernel, which leads to the conclusion. \square

From there follows directly the following result.

Lemma 17. *For every design set with coordinate-distinct points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$, there exists a constant $c_{\mathbf{x}} > 0$ such that whatever $\boldsymbol{\theta} \in (\mathbb{R}_+^*)^r$,*

$$\forall \boldsymbol{\xi} = (\xi_1, \dots, \xi_n) \in \mathbb{R}^n, \sum_{j,k=1}^n \xi_j \xi_k K_{r,\nu} \left(\frac{\mathbf{x}^{(j)} - \mathbf{x}^{(k)}}{\boldsymbol{\theta}} \right) \geq c_{\mathbf{x}} \|\boldsymbol{\xi}\|^2 2^{-\frac{r}{2}-\nu} M_r(\nu) f_{r,\nu}(\theta_r) \quad (40)$$

where $f_{r,\nu}(\theta_r) = (2\sqrt{\nu})^{-r-2\nu} \theta_r^r$ if $\theta_r \leq 2\sqrt{\nu}$ and $f_{r,\nu}(\theta_r) = \theta_r^{-2\nu}$ if $\theta_r \geq 2\sqrt{\nu}$.

Proof. For every design set $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$, the set of all design sets that can be written $\frac{\theta_r}{\theta} \mathbf{x}^{(1)}, \dots, \frac{\theta_r}{\theta} \mathbf{x}^{(n)}$ ($\theta \in (\mathbb{R}_+^*)^r$) has compact closure. If the design set $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ has coordinate-distinct points, then every design set in the aforementioned closure has no overlapping points. Thus the conclusion follows from lemma 16. \square

Lemma 18. *There exists a correlation kernel $\tilde{K}_{r,\nu}^{tens}$ such that, whatever the design set $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$, whatever $\theta \in (\mathbb{R}_+^*)^r$ and $\xi = (\xi_1, \dots, \xi_n) \in \mathbb{R}^n$,*

$$\sum_{j,k=1}^n \xi_j \xi_k K_{r,\nu}^{tens} \left(\frac{\mathbf{x}^{(j)} - \mathbf{x}^{(k)}}{\theta} \right) \geq 2^{-\frac{r}{2}-\nu} M(\nu)^r f_{r,\nu}(\theta_r) \sum_{j,k=1}^n \xi_j \xi_k \tilde{K}_{r,\nu}^{tens} \left(\frac{\theta_r}{\theta} (\mathbf{x}^{(j)} - \mathbf{x}^{(k)}) \right) \quad (41)$$

where $f_{r,\nu}(\theta) = (2\sqrt{\nu})^{-r-2\nu} \theta^r$ if $\theta \leq 2\sqrt{\nu}$ and $f_{r,\nu}(\theta) = \theta^{-2\nu}$ if $\theta \geq 2\sqrt{\nu}$.

Proof. Whatever $\mathbf{x}, \mathbf{y} \in \mathbb{R}^r$, $K_{r,\nu}^{tens}(\mathbf{x} - \mathbf{y}) = \int_{\mathbb{R}^r} \hat{K}_{r,\nu}^{tens}(\omega) e^{i\langle \omega | \mathbf{x} - \mathbf{y} \rangle} d\omega = \int_{\mathbb{R}^r} \prod_{k=1}^r \hat{K}_\nu(\omega_k) e^{i\langle \omega | \mathbf{x} - \mathbf{y} \rangle} d\omega$. Let us define $I_1^r := \{\mathbf{s} \in \mathbb{R}^r : \forall 1 \leq k \leq r, |\mathbf{s}_k| \geq 1\}$.

$$\begin{aligned} \sum_{j,k=1}^n \xi_j \xi_k K_{r,\nu}^{tens} \left(\frac{\mathbf{x}^{(j)} - \mathbf{x}^{(k)}}{\theta} \right) &= \int_{\mathbb{R}^r} \hat{K}_{r,\nu}^{tens}(\omega) \left| \sum_{j=1}^n \xi_j e^{i\langle \omega | \frac{\mathbf{x}^{(j)}}{\theta} \rangle} \right|^2 d\omega \\ &= M(\nu)^r \underline{\theta}_r^r \int_{\mathbb{R}^r} \prod_{k=1}^r (4\nu + \underline{\theta}_r^2 \mathbf{s}_k^2)^{-\frac{1}{2}-\nu} \left| \sum_{j=1}^n \xi_j e^{i\langle \mathbf{s} | \frac{\theta_r}{\theta} \mathbf{x}^{(j)} \rangle} \right|^2 d\mathbf{s} \\ &\geq 2^{-\frac{r}{2}-\nu} M(\nu)^r f_{r,\nu}(\theta_r) \int_{I_1^r} \prod_{k=1}^r |\mathbf{s}_k|^{-1-2\nu} \left| \sum_{j=1}^n \xi_j e^{i\langle \mathbf{s} | \frac{\theta_r}{\theta} \mathbf{x}^{(j)} \rangle} \right|^2 d\mathbf{s} \end{aligned} \quad (42)$$

Now, let $\tilde{K}_{r,\nu}^{tens}$ be the function with Fourier transform $\hat{\tilde{K}}_{r,\nu}^{tens}(\omega) = \prod_{k=1}^r \mathbf{1}_{\{|\omega_k| \geq 1\}} \omega_k^{-1-2\nu}$. According to Bochner's theorem, $\tilde{K}_{r,\nu}^{tens}$ is a correlation kernel, which leads to the conclusion. \square

From there follows directly the following result.

Lemma 19. *For every design set with coordinate-distinct points $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$, there exists a constant $c_{\mathbf{x}}^{tens} > 0$ such that whatever $\theta \in (\mathbb{R}_+^*)^r$,*

$$\forall \xi = (\xi_1, \dots, \xi_n) \in \mathbb{R}^n, \sum_{j,k=1}^n \xi_j \xi_k K_{r,\nu}^{tens} \left(\frac{\mathbf{x}^{(j)} - \mathbf{x}^{(k)}}{\theta} \right) \geq c_{\mathbf{x}}^{tens} \|\xi\|^2 2^{-\frac{r}{2}-\nu} M_r(\nu) f_{r,\nu}(\theta_r) \quad (43)$$

where $f_{r,\nu}(\theta_r) = (2\sqrt{\nu})^{-r-2\nu} \theta_r^r$ if $\theta_r \leq 2\sqrt{\nu}$ and $f_{r,\nu}(\theta_r) = \theta_r^{-2\nu}$ if $\theta_r \geq 2\sqrt{\nu}$.

Proof. For every design set $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$, the set of all design sets that can be written $\frac{\theta_r}{\theta} \mathbf{x}^{(1)}, \dots, \frac{\theta_r}{\theta} \mathbf{x}^{(n)}$ ($\theta \in (\mathbb{R}_+^*)^r$) has compact closure. If the design set $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ has coordinate-distinct points, then every design set in the aforementioned closure has no overlapping points. Thus the conclusion follows from lemma 18. \square

Proposition 20. *Majoration 6 of assumption 2 with $a \geq 2\nu$ is true for Matérn geometric anisotropic and tensorized kernels as long as the design set has coordinate-distinct points.*

Proof. Let us use the norm $\|\mathbf{M}\|_\infty = \sup\{\|\mathbf{M}\mathbf{x}\|, \|\mathbf{x}\| = 1\}$. If \mathbf{M} is diagonalizable, this norm is equal to \mathbf{M} 's eigenvalue with greatest absolute value. Thus, the norm of \mathbf{M}^{-1} is the inverse of \mathbf{M} 's eigenvalue with smallest absolute value. Thus lemmas 17 and 19 prove majoration 6 if we replace $\boldsymbol{\xi}$ with a unit eigenvector corresponding to $\boldsymbol{\Sigma}_\theta$'s smallest eigenvalue (as $\boldsymbol{\Sigma}_\theta$ is positive definite, all eigenvalues are positive). That we chose $\|\cdot\|_\infty$ as our norm is of no concern, as all norms are equivalent in finite-dimensional vectorial spaces. \square

Proof of Proposition 5. Combining Proposition 15 and Proposition 20 yields all majorations but majoration 5 for any $a \geq 2\nu$ and any $1 < b \leq 1 + 2\min(1, \nu)$ if $\nu \neq 1$ (resp. $a \geq 2$ and $1 < b < 2$ if $\nu = 1$). Now, majoration 5 is obviously true for any a and b . Thus we get the desired result. \square

Acknowledgements

The author would like to thank his PhD advisor Professor Josselin Garnier (École Polytechnique, Centre de Mathématiques Appliquées) for his guidance, Loic Le Gratiet (EDF R&D, Chatou) and Anne Dutfoy (EDF R&D, Saclay) for their advice and helpful suggestions. The author would also like to thank his colleagues at EDF for the many insightful discussions which helped further this work, and Électricité de France for its financial support.

References

- M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *Applied Mathematics Series*. National Bureau of Standards, 1964.
- E. Anderes. On the consistent separation of scale and variance for Gaussian random fields. *Annals of Statistics*, 38:870–893, 2010.
- I. Andrianakis and P. G. Challenor. The effect of the nugget on gaussian process emulators of computer models. *Computational Statistics & Data Analysis*, 56(12):4215–4228, 2012.
- F. Bachoc. *Parametric estimation of covariance function in Gaussian-process based Kriging models. Application to uncertainty quantification for computer models*. PhD thesis, Université Paris Diderot, 2013.
- F. Bachoc. Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes. *Journal of Multivariate Analysis*, 125:1–35, 2014.
- J. O. Berger and J. M. Bernardo. On the Development of Reference Priors. *Bayesian statistics*, 4(4):35–60, 1992.
- J. O. Berger, V. De Oliveira, and B. Sansó. Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456):1361–1374, 2001.
- J. O. Berger, J. M. Bernardo, and D. Sun. The formal definition of reference priors. *Annals of Statistics*, 37(2):905–938, 2009.
- J. M. Bernardo. Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 113–147, 1979.
- B. S. Clarke and A. R. Barron. Jeffreys’ prior is asymptotically least favorable under entropy risk. *Journal of Statistical planning and Inference*, 41(1):37–60, 1994.
- V. De Oliveira, B. Kedem, and D. A. Short. Bayesian Prediction of Transformed Gaussian Random Fields. *Journal of the American Statistical Association*, 92:1422–1433, 1997.
- J.C. Gower. Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra and its Applications*, 67:81–97, 1985.
- M. Gu. *Robust Uncertainty Quantification and Scalable Computation for Computer Models with Massive Output*. PhD thesis, Duke University, 2016.
- M. S. Handcock and M. L. Stein. A Bayesian Analysis of Kriging. *Technometrics*, 35:403–410, 1993.
- H. Hotelling. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2):193–232, 1953.
- A. G. Journel and Ch. J. Huijbregts. *Mining geostatistics*. Academic press, 1978.
- H. Kazianka and J. Pilz. Objective bayesian analysis of spatial data with uncertain nugget and range parameters. *Canadian Journal of Statistics*, 40(2):304–327, 2012.

- M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.
- R. Li and A. Sudjianto. Analysis of computer experiments using penalized likelihood in gaussian kriging models. *Technometrics*, 47(2):111–120, 2005.
- B. Matérn. *Spatial Variation*. Springer-Verlag, Berlin, 2nd edition, 1986.
- R. Paulo. Default priors for Gaussian processes. *Annals of Statistics*, 33(2):556–582, 2005.
- C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.
- C. Ren, D. Sun, and C. He. Objective bayesian analysis for a spatial model with nugget effects. *Journal of Statistical Planning and Inference*, 142(7):1933–1946, 2012.
- C. Ren, D. Sun, and S. K. Sahu. Objective bayesian analysis of spatial models with separable correlation functions. *Canadian Journal of Statistics*, 41(3):488–507, 2013.
- C. P. Robert, N. Chopin, and J. Rousseau. Harold Jeffreys's Theory of Probability Revisited. *Statistical Science*, 24(2):141–172, 2009.
- T. J. Santner, B. J. Williams, and W. I. Notz. *The Design and Analysis of Computer Experiments*. Springer-Verlag, New York, 2003.
- I.T. Schoenberg. On certain Metric Spaces arising from Euclidean Spaces by a change of metric and their imbedding in Hilbert Space. *Annals of Mathematics*, 38(4):787–793, 1937.
- M. L. Stein. *Interpolation of Spatial Data. Some Theory for Kriging*. Springer Series in Statistics. Springer-Verlag, New York, 1999.
- H. Wackernagel. *Multivariate Geostatistics*. Springer-Verlag, Berlin, 1995.
- A. M. Yaglom. *Correlation Theory of Stationary and Related Random Functions 1. Basic Results*. Springer-Verlag, New York, 1987.
- H. Zhang. Inconsistent Estimation and Asymptotically Equal Interpolations in Model-based Geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004.